# LOGISTISCHE REGRESSIE

4 juni 2024

Training O + S

Elmar Jansen (elmar@elmarjansen.nl)

# VANDAAG

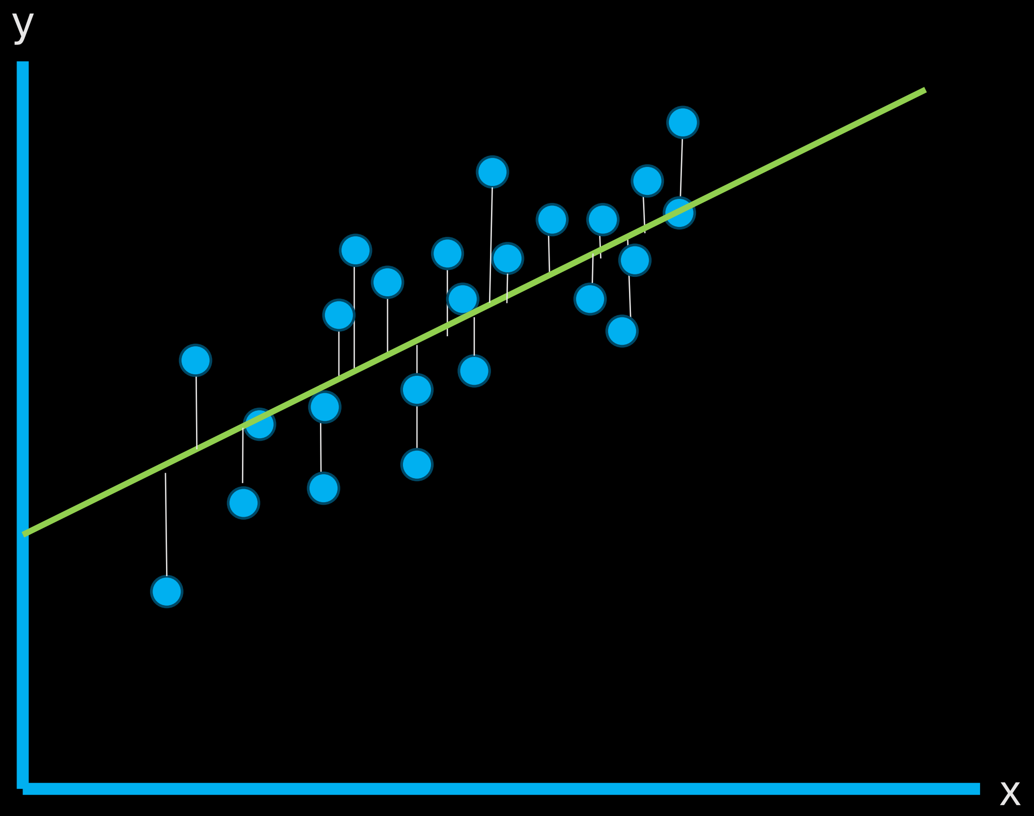
1. Terugblik
2. Logistische Regressie: waarom?
3. De Logit-transformatie
4. Coefficienten interpreteren
5. Pseudo R2

# DE KOMENDE WEKEN

| Bijeenkomst | Onderwerp |
|---|---|
| Dinsdag 14 mei | Lineaire regressie: de basis |
| Dinsdag 21 mei | Lineaire regressie vervolg: assumpties en controleren |
| Donderdag 30 mei | Interacties en dummy-variabelen |
| Dinsdag 4 juni | Logistische Regressie |
| Dinsdag 11 juni | Multilevel-analyse |

# TERUGBLIK

# LINEAIRE REGRESSIE




$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

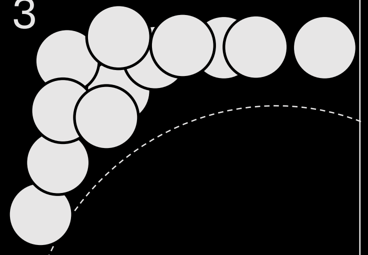
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \varepsilon_i$$

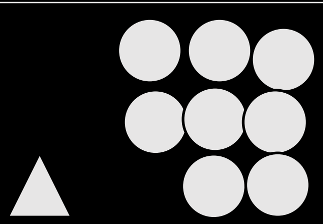
# 8 GEVAREN VAN REGRESSIE

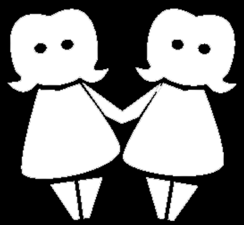
DANGER!!

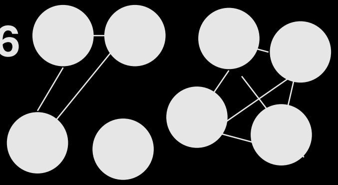
1 
Schijnverband

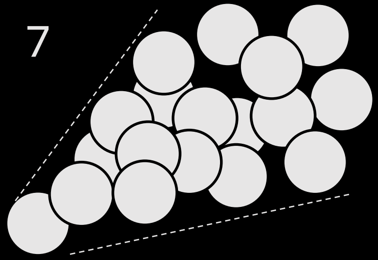
2 
Wederkerigheid / Simultaniteit

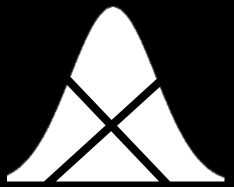
3 
Non-Lineairiteit

4 
Extreme waarden

5 
Multicollineariteit

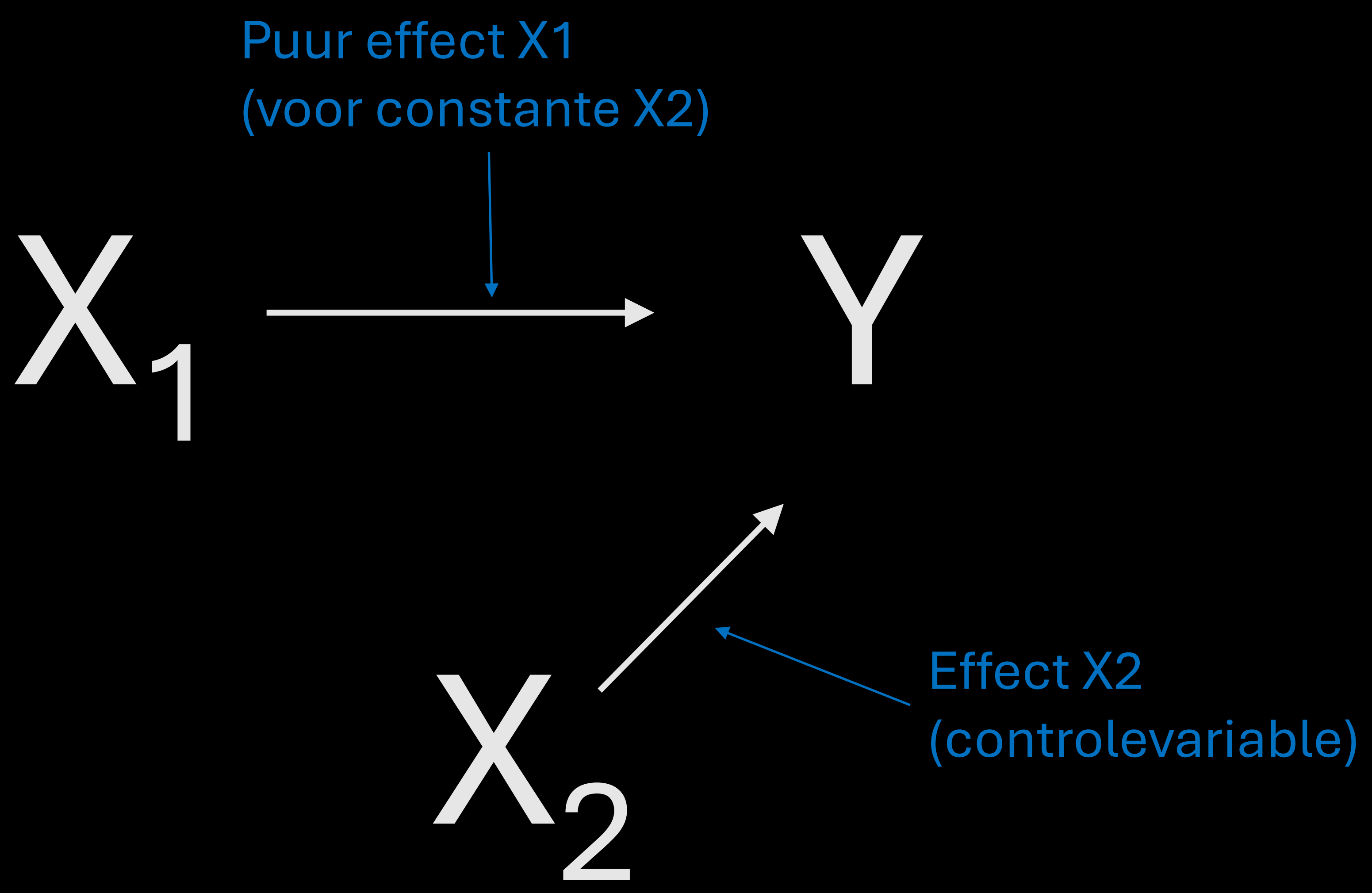
6 
Niet onafhankelijke residuen

7 
Heteroskedasticiteit

8 
Non-normaliteit van residuen

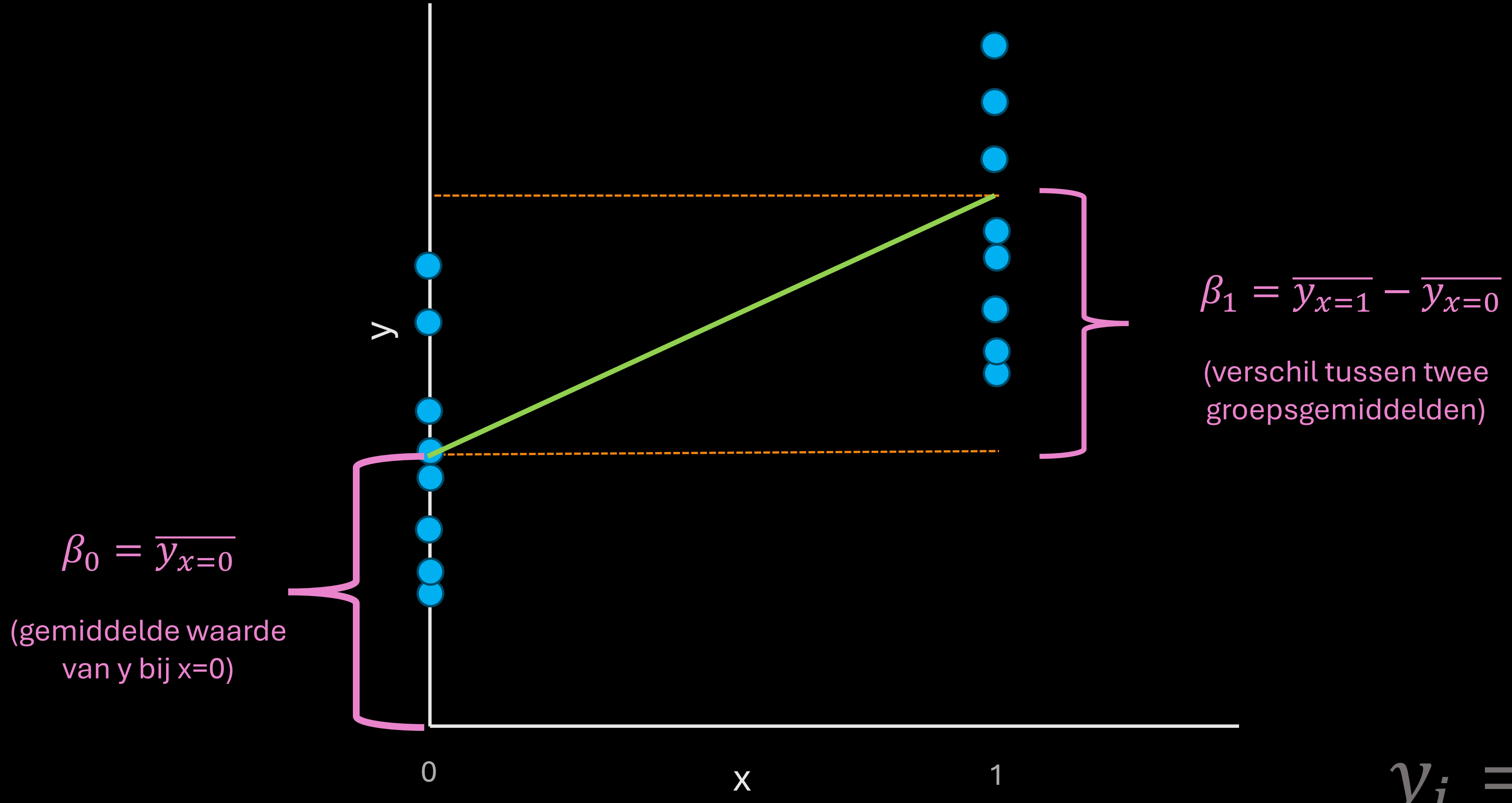
# CONTROLEREN

Door onafhankelijke variabele $X_2$ toe te voegen aan het model krijgen we het effect van $X_1$ **constant houdend voor $X_2$** (en viceversa)



# DUMMY-VARIABELE

**Dichotome variabele met waarden 0 en 1 als onafhankelijke variabele**




$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

# DUMMY-VARIABELE

**We maken een dichotome variabele met waarden 0 en 1**

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

**X = 0** $\quad y_i = \beta_0 + \beta_1 \times 0 + \varepsilon_i = \boxed{\beta_0 + \varepsilon_i}$

$\beta_0 = \overline{y_{x=0}}$

(gemiddelde waarde van y bij x=0)

---

**X = 1** $\quad y_i = \beta_0 + \beta_1 \times 1 + \varepsilon_i = \boxed{\beta_0 + \beta_1 + \varepsilon_i}$

$\beta_1 = \overline{y_{x=1}} - \overline{y_{x=0}}$

(verschil tussen twee groepsgemiddelden)

# CATEGORIALE ONAFHANKELIJKE VARIABELEN

| Categorie A | Categorie B | Categorie C |
|---|---|---|
| $CatB_i = 0$<br>$CatC_i = 0$ | $CatB_i = 1$<br>$CatC_i = 0$ | $CatB_i = 0$<br>$CatC_i = 1$ |

Baseline-categorie

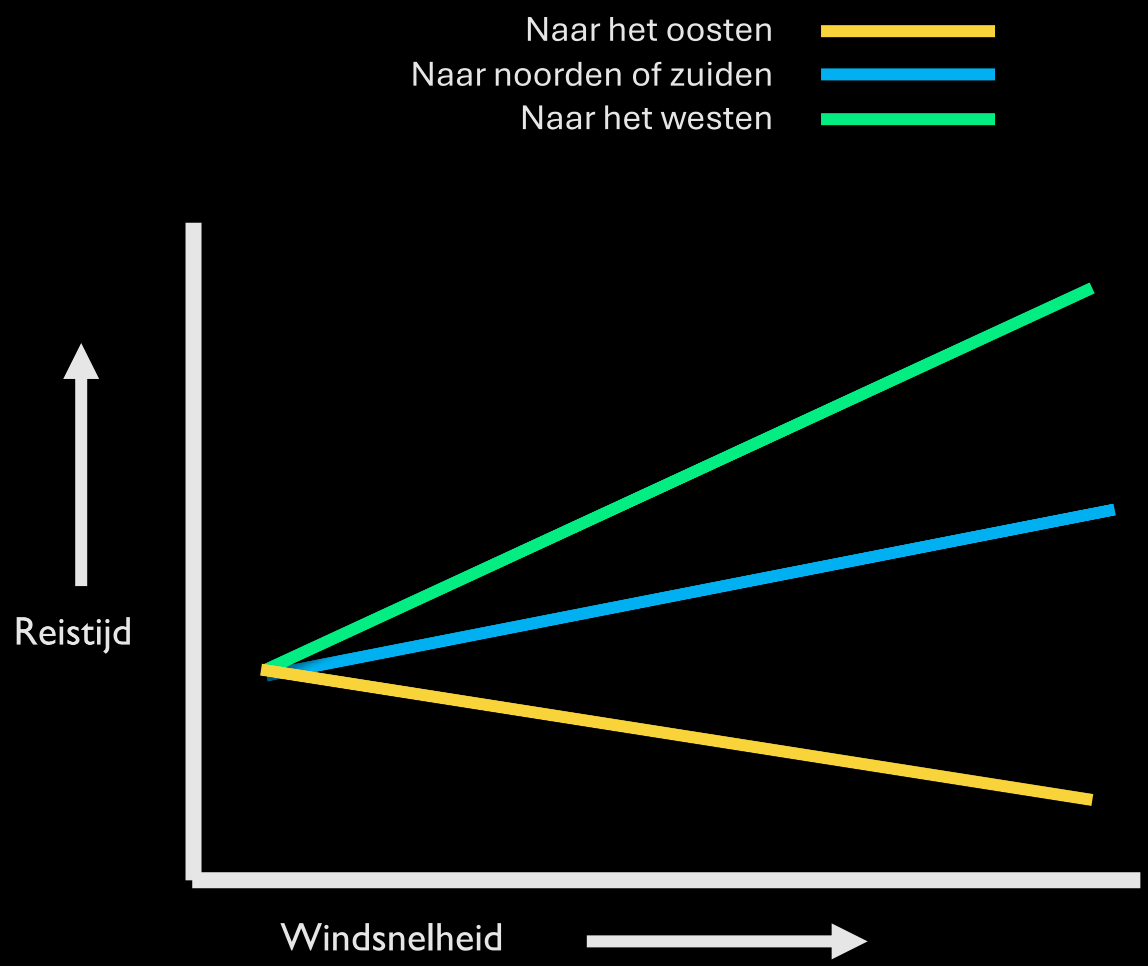
# INTERACTIE-EFFECT

Een effect

van een variabele

*op het effect van*

een andere variabele



# INTERACTIE-EFFECT IN REGRESSIE

Voeg ook altijd het "main"-effect van beide variabelen toe!

De interactie is de vermenigvuldiging tussen beide variabelen

$$Y_i = \beta_0 + \boxed{\beta_1 X_{1i} + \beta_2 X_{2i}} + \boxed{\beta_3 X_{1i} X_{2i}} + \varepsilon_i$$

Als je herschikt, zie je dat het effect van X1 nu afhankelijk is van X2:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

$$Y_i = \boxed{(\beta_0 + \beta_2 X_{2i})} + \boxed{(\beta_1 + \beta_3 X_{2i})} X_{1i} + \varepsilon_i$$

Effect van X1 wordt nu zelf beïnvloed door X2

# INTERPRETATIE INTERACTIE

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$$

---

$\beta_0$: Waarde $Y_i$ *als* $X_{1i}$ *en* $X_{2i}$ *0 zijn*

$\beta_1$: Effect van $X_{1i}$ *als* $X_{2i}$ *0 is* (dus: hoeveel stijgt $Y_i$ als $X_{1i}$ met 1 omhoog gaat als $X_{2i}$ 0 is)

$\beta_2$: Effect van $X_{2i}$ *als* $X_{1i}$*0 is* (dus: hoeveel stijgt $Y_i$ als $X_{2i}$ met 1 omhoog gaat als $X_{1i}$ 0 is)

$\beta_3$: Effect van $X_{2i}$ op effect van $X_{1i}$ (of effect van $X_{1i}$ op effect van $X_{2i}$)

*(dus met hoeveel meer stijgt $Y_i$ voor iedere toename van $X_{1i}$ met 1 wanneer $X_{2i}$ met 1 omhoog gaat)*

*of*

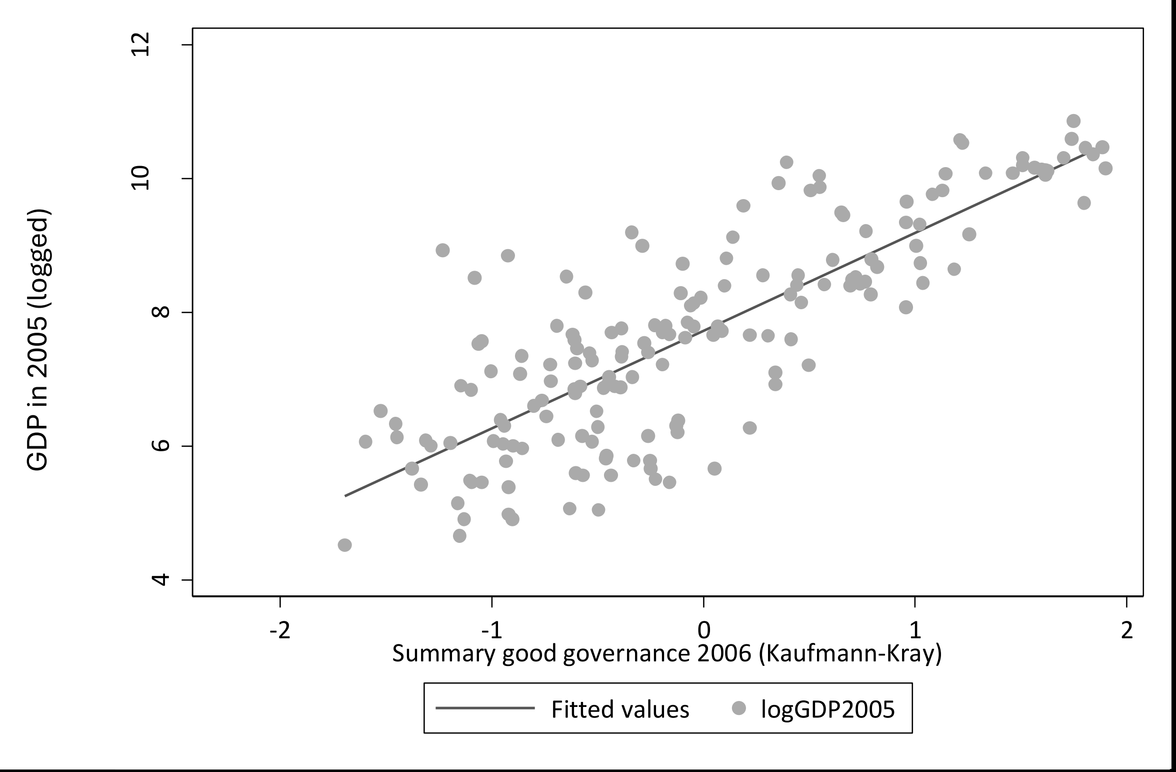
*(dus met hoeveel meer stijgt $Y_i$ voor iedere toename van $X_{2i}$ met 1 wanneer $X_{1i}$ met 1 omhoog gaat)*

# DUMMY ALS AFHANKELIJKE VARIABELEN

Logistische regressie:
waarom?

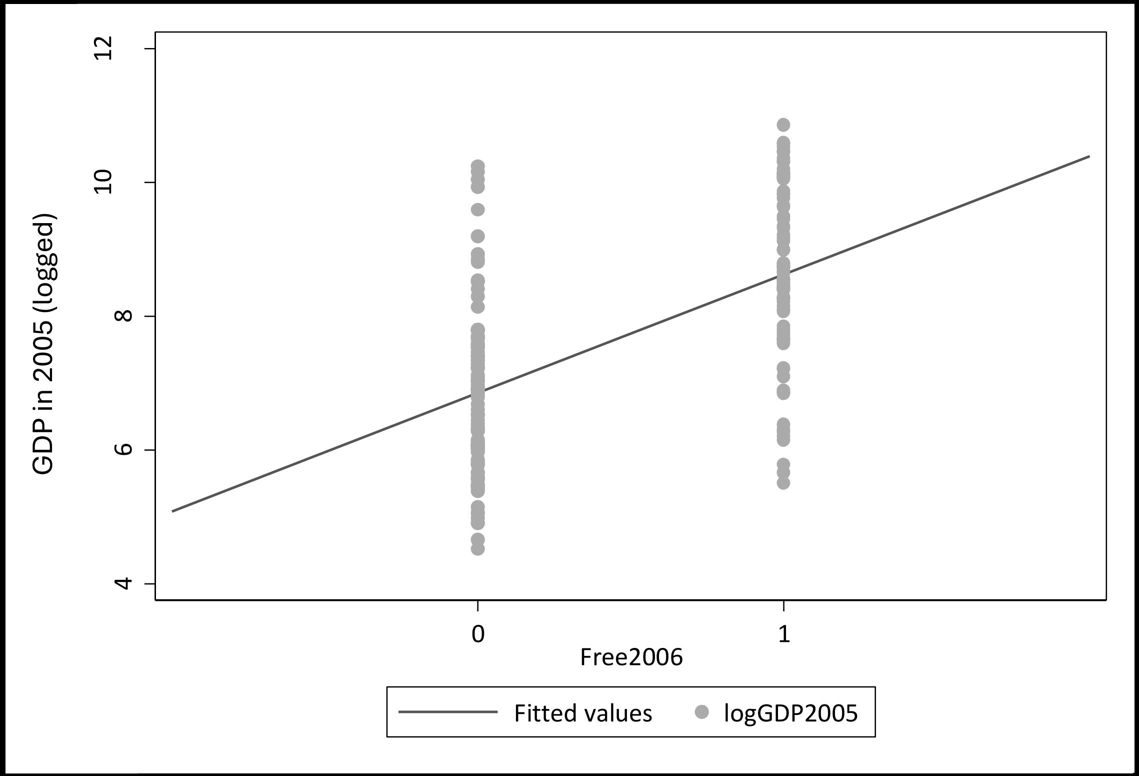
# LINEAIRE REGRESSIE (OLS)

Tot nu toe gebruikten we *lineaire* regressie om een effect te schatten als een *rechte lijn*



# LINEAIRE REGRESSIE (OLS)

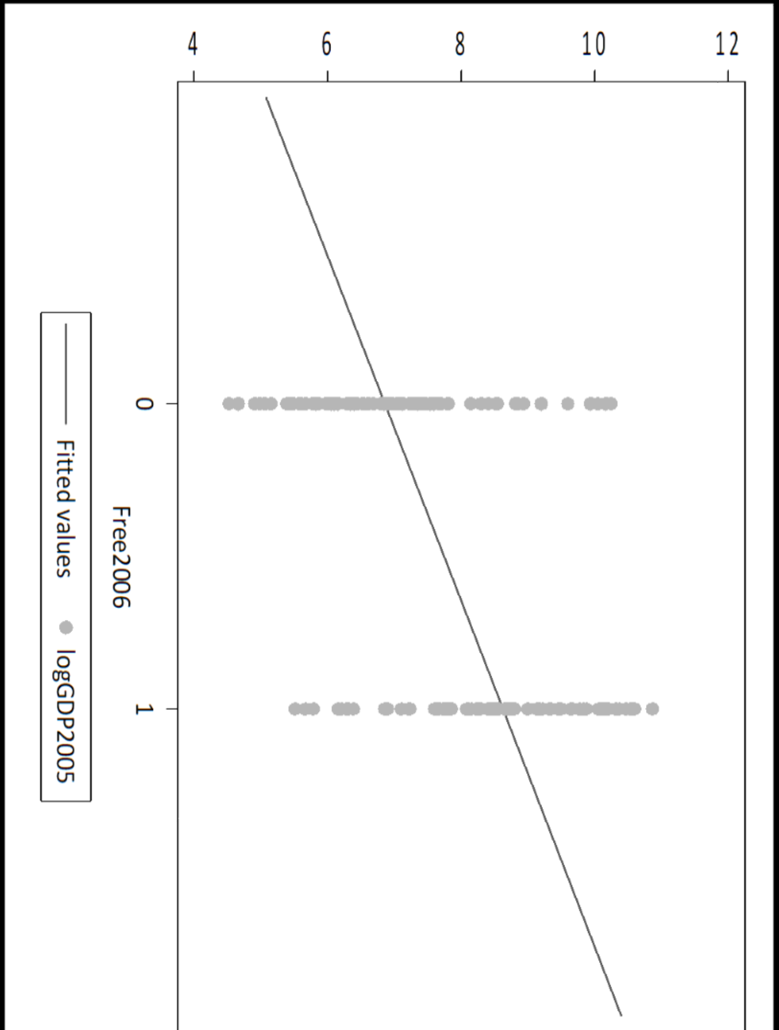
Tot nu toe gebruikten we *lineaire* regressie om een effect te schatten als een *rechte lijn*




Dichtome **<u>on</u>**afhankelijke met een dummy

# LINEAR REGRESSION (OLS)

Tot nu toe gebruikten we *lineaire* regressie om een effect te schatten als een *rechte lijn*




*Maar wat als we de X-as en Y-as verwisselen?*

Wat als we iets willen *voorspellen* dat

niet *continu maar dichotoom is?*


-Stemmen of niet stemmen

-Een huis bezitten of geen huis bezitten

-Een bepaald symptoom wel of niet hebben


-Je rijbewijs halen of niet halen

-Wel of niet een strike gooien

# EENVOUDIG VOORBEELD

Afhankelijke Variabele

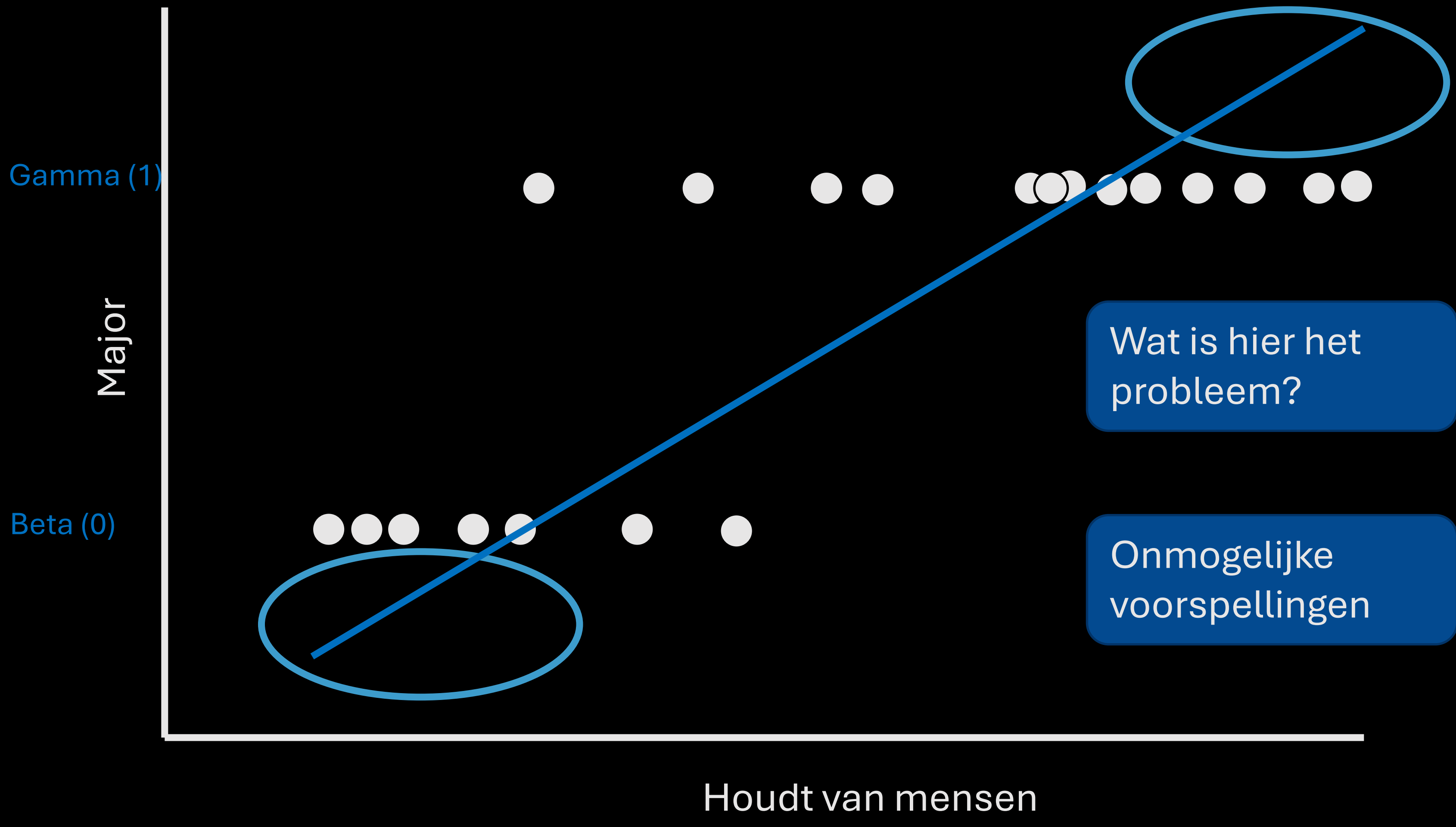
*Keuze voor een gamma-major*

Onafhankelijke variabele

*Houdt student van mensen?*

# REGRESSIE MET DUMMY ALS AFHANKELIJKE

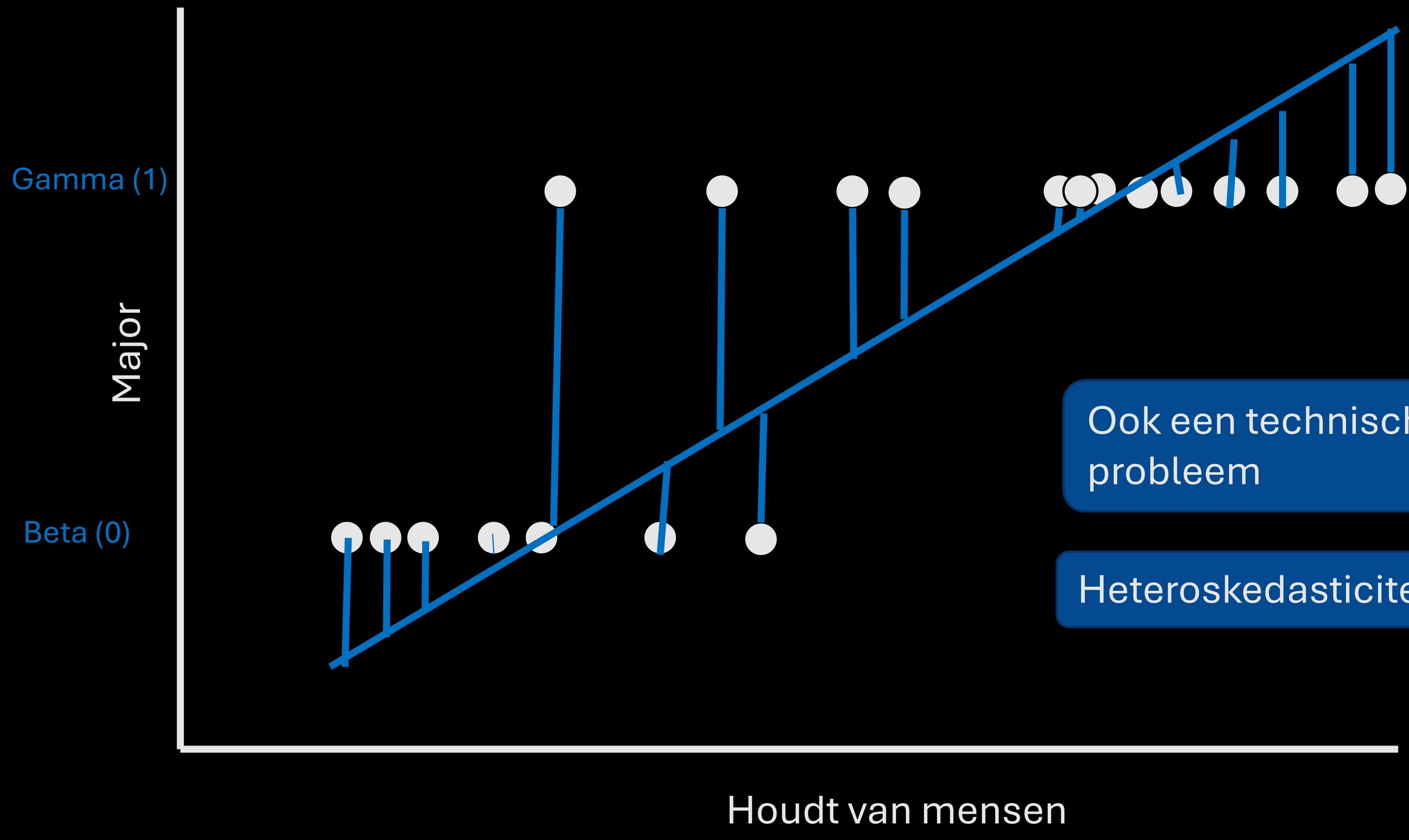
$$gamma\_major_i = 0{,}30 + 0{,}15 \times mensen_i + e_i$$




Wat is hier het probleem?

Onmogelijke voorspellingen

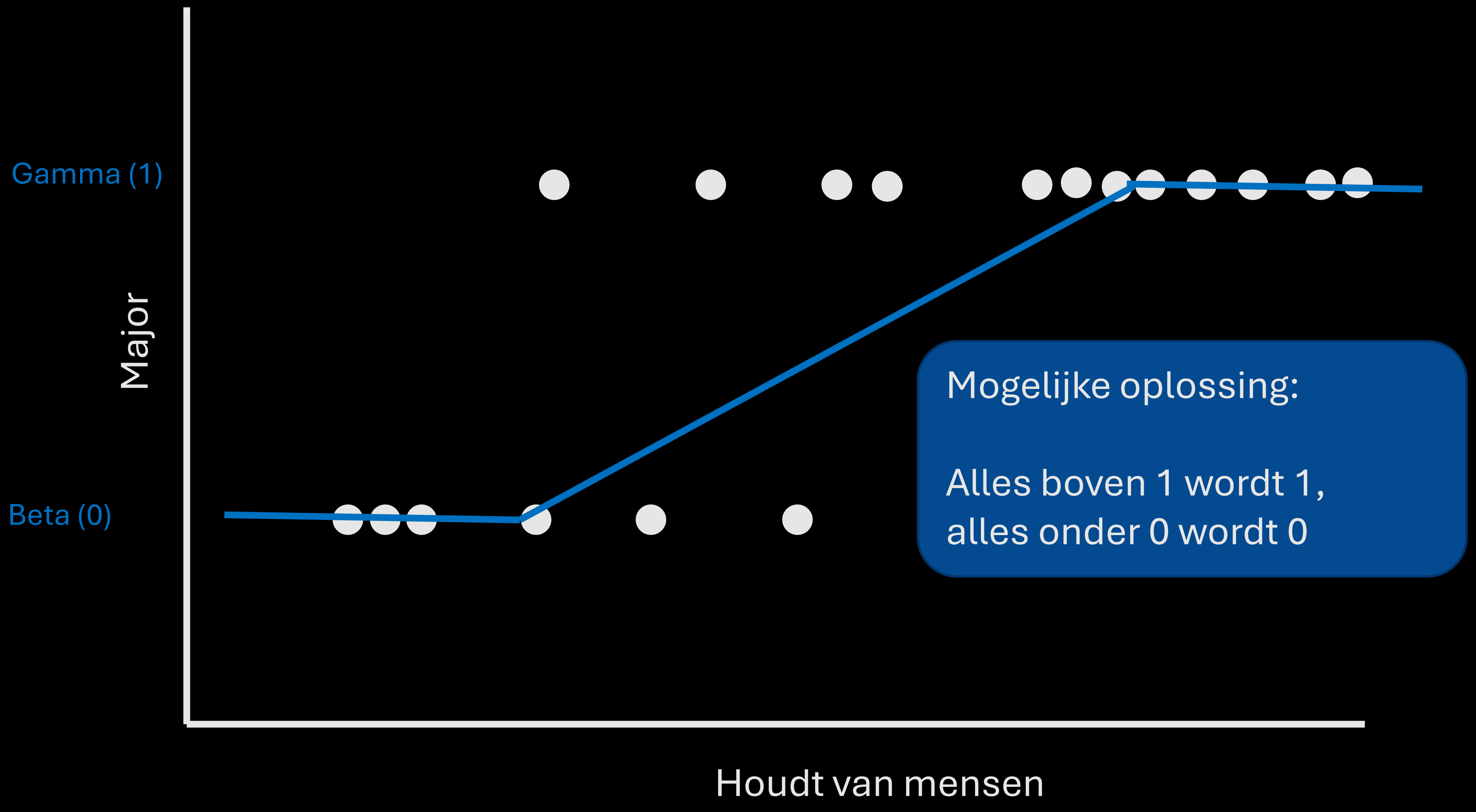
# REGRESSIE MET DUMMY ALS AFHANKELIJKE




Ook een technisch probleem

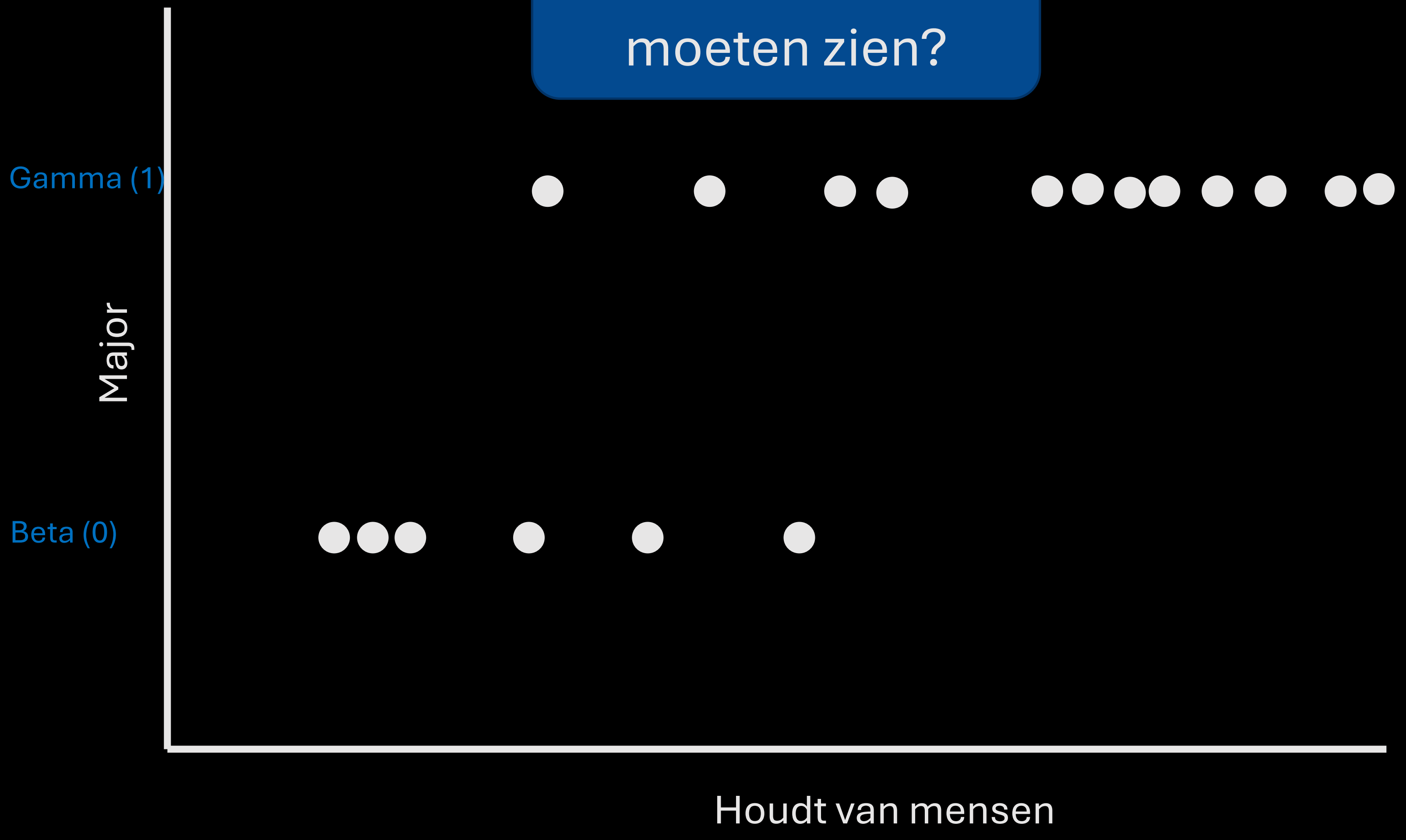
Heteroskedasticiteit

# REGRESSIE MET DUMMY ALS AFHANKELIJKE

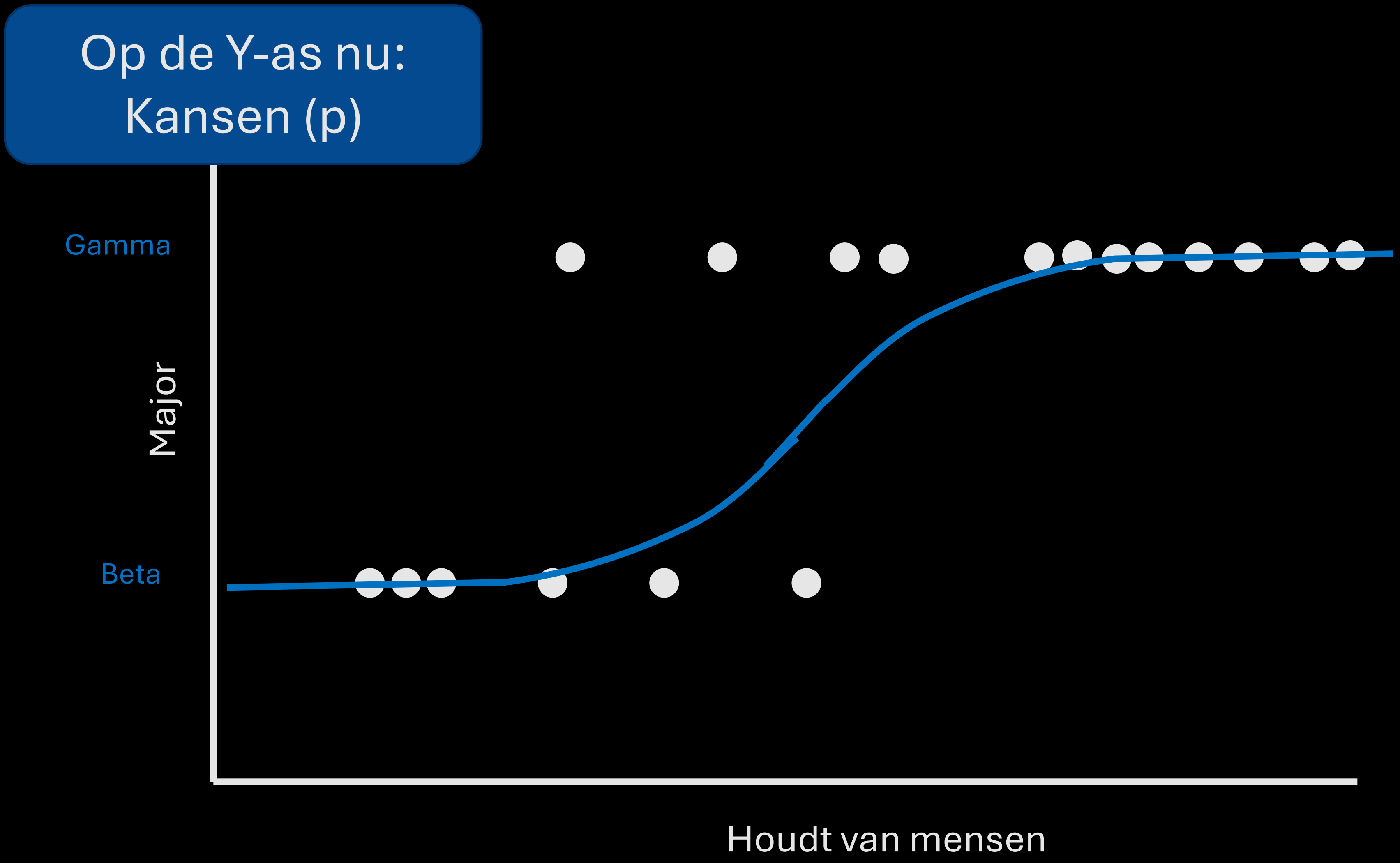


# REGRESSIE MET DUMMY ALS AFHANKELIJKE

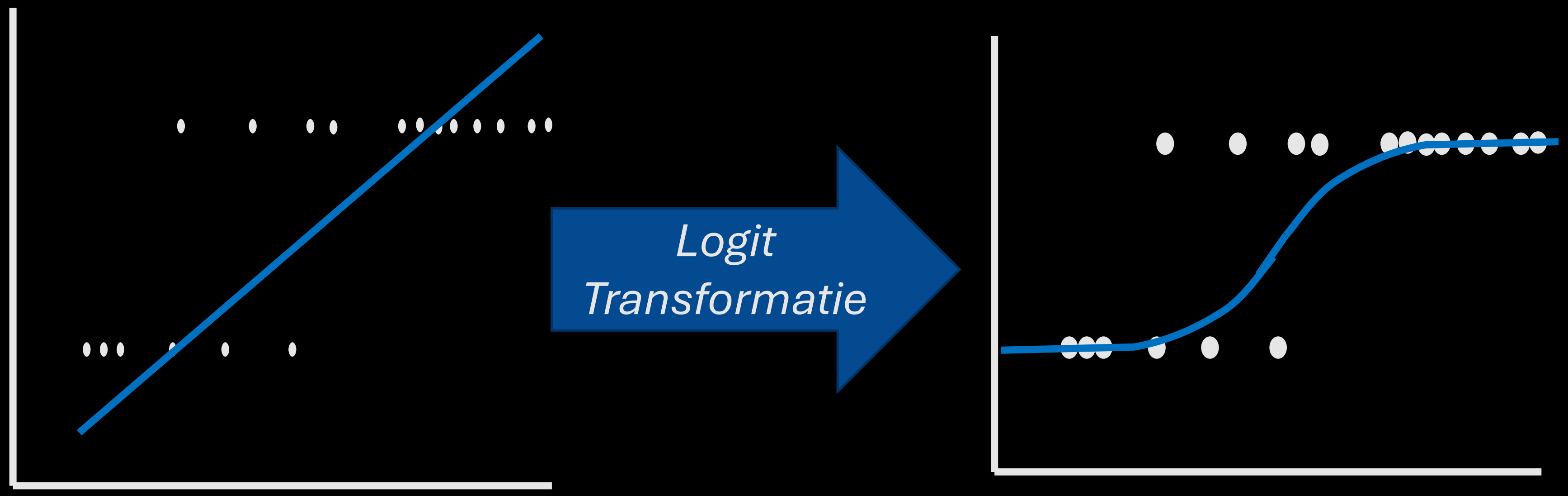
Hoe zou het er uit moeten zien?



# REGRESSIE MET DUMMY ALS AFHANKELIJKE



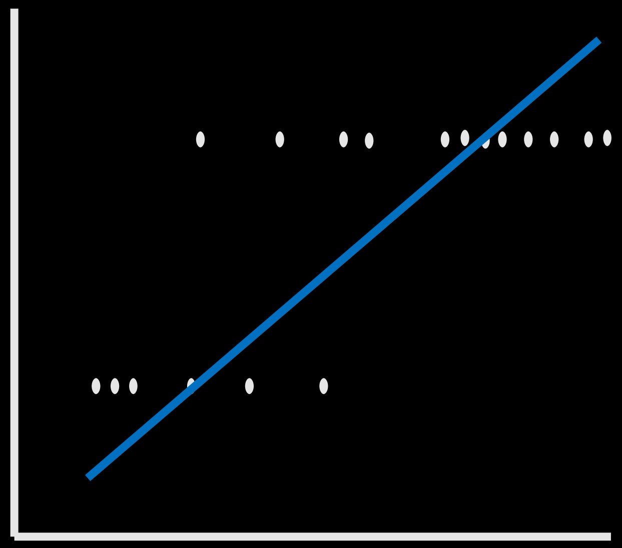
# HOE KOMEN WE DAAR?



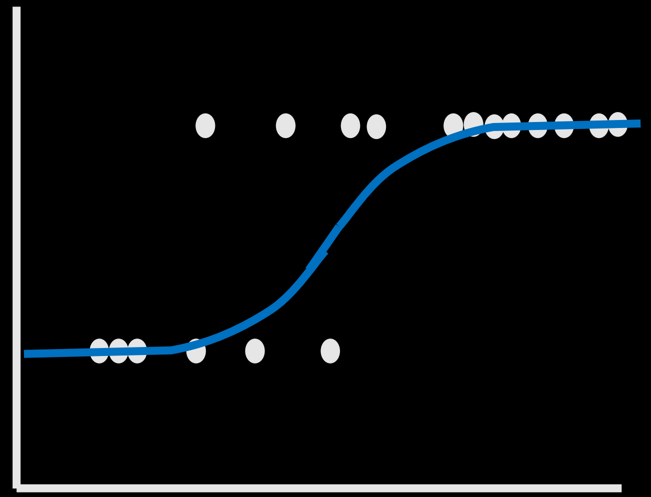

$$\boxed{y_i} = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\boxed{\ln \frac{P_i}{1 - P_i}} = \beta_0 + \beta_1 X_i$$

# LOGIT TRANSFORMATIE



$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



$$\ln \frac{P_i}{1 - P_i} = \beta_0 + \beta_1 X_i$$

Beginnend bij kans P,
wat moet er gebeuren?

# LOGIT TRANSFORMATIE

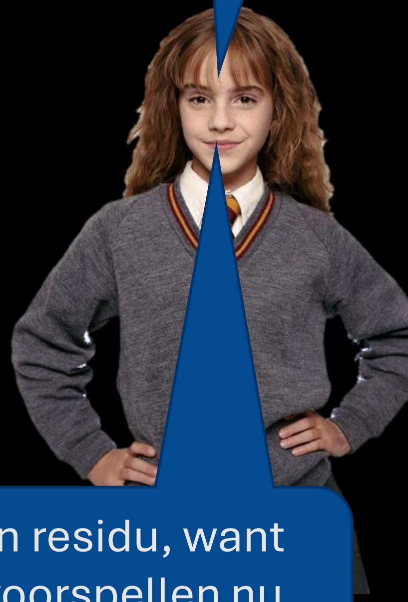
## Stap 1:

gebruik *odds* in plaats van kansen  om plafond 1 weg te werken

## Stap 2:

neem natuurlijk logaritme om bodem 0 weg te werken

Logit-transformative om kansen om te zetten in (lineaire) logged odds



Geen residu, want we voorspellen nu *kansen*.

Lineair deel

<u>Logged odds</u>

= ‘logit’ transformatie

$$\ln \frac{P_i}{1 - P_i} = \boxed{\beta_0 + \beta_1 X_i}$$

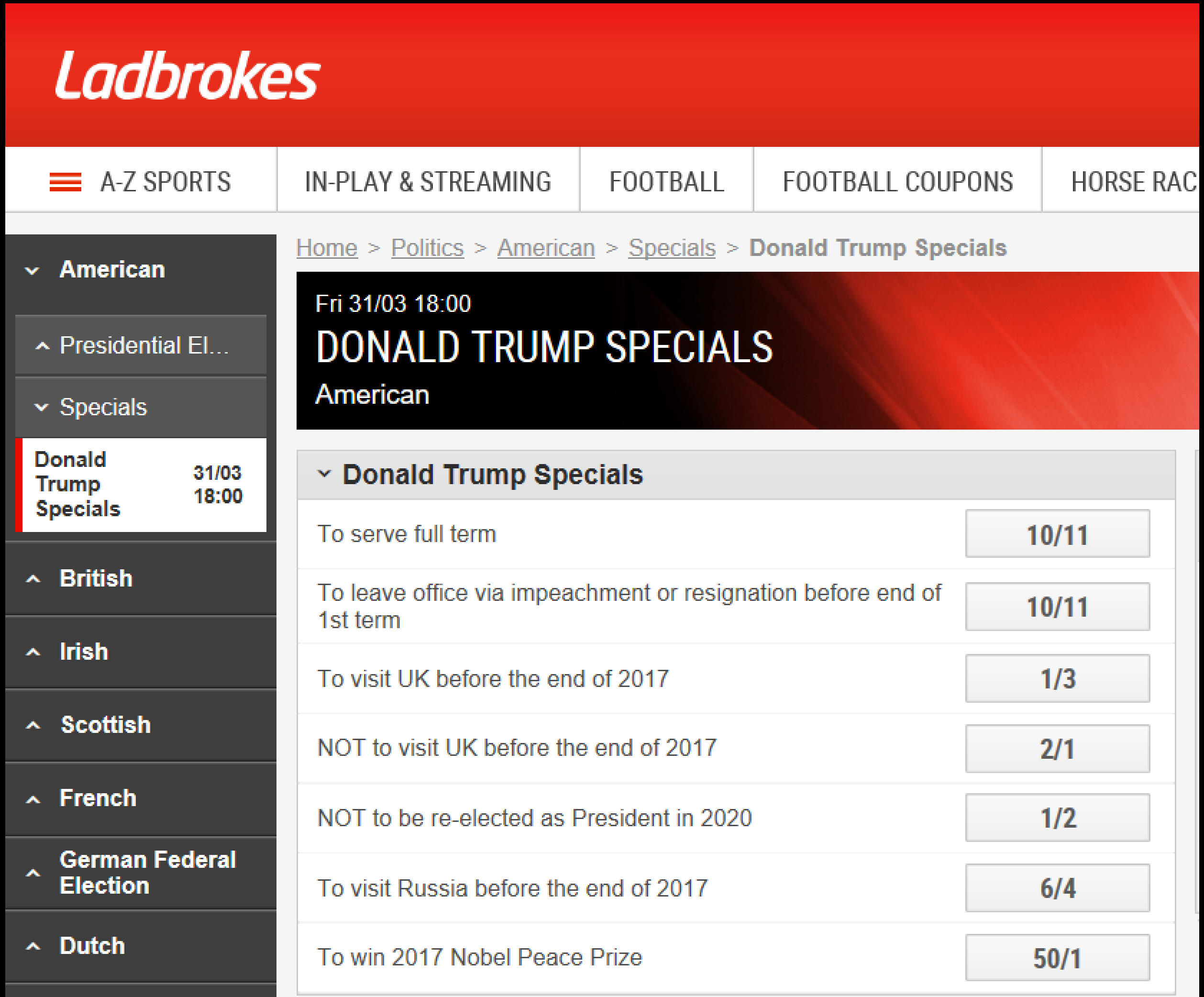
Natuurlijk logaritme

<u>Odds</u>: kans dat iets gebeurt, gedeeld door kans dat het niet gebeurt

# STAP 1: ODDS

Logit transformatie in twee stappen

# ODDS


Ladbrokes
A-Z SPORTS
IN-PLAY & STREAMING
FOOTBALL
FOOTBALL COUPONS
HORSE RAC
American
Presidential El…
Specials
Donald Trump Specials
31/03 18:00
British
Irish
Scottish
French
German Federal Election
Dutch
Home > Politics > American > Specials > Donald Trump Specials
Fri 31/03 18:00
DONALD TRUMP SPECIALS
American
Donald Trump Specials
To serve full term 10/11
To leave office via impeachment or resignation before end of 1st term 10/11
To visit UK before the end of 2017 1/3
NOT to visit UK before the end of 2017 2/1
NOT to be re-elected as President in 2020 1/2
To visit Russia before the end of 2017 6/4
To win 2017 Nobel Peace Prize 50/1

# ODDS EN KANSEN

Kans (of aantal keer) dat iets gebeurt,
gedeeld door kans (of aantal keer) dat het niet gebeurt.

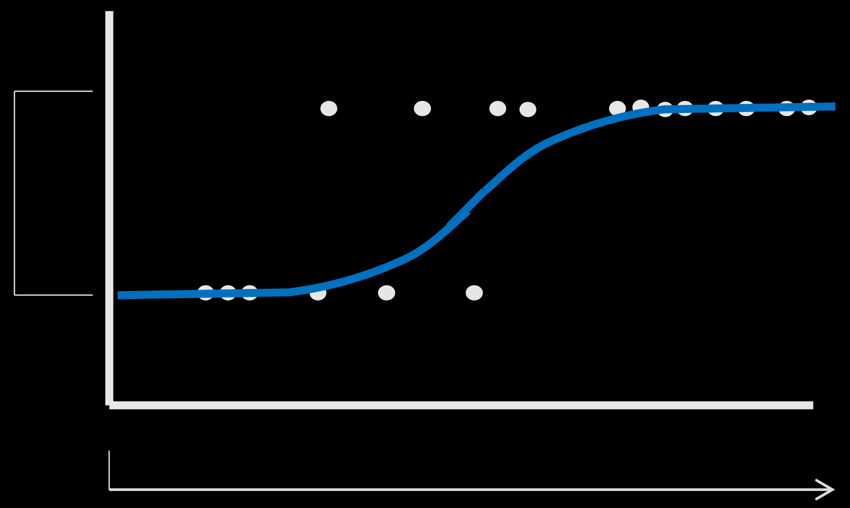
(P = kans)

| | | |
|---|---|---|
| P | => | Odds = P / (1 – P) |
| P = 0,5 | => | Odds = 1 (gelijke kans) |
| P = 0,8 | => | Odds = 4 (gebeurt 4x voor iedere 1x niet) |
| P = 0,2 | => | Odds = ¼ (gebeurt ¼x voor iedere 1x niet) |

| | | |
|---|---|---|
| P < ½ | => | Odds tussen 0 en 1 |
| P > ½ | => | Odds tussen 1 en oneindig |

# WAAROM KANSEN OMZETTEN IN ODDS?

- Kansen vallen tussen 0 en 1
- Odds vallen tussen 0 en oneindig



Onze lineaire voorspelling kan tot oneindig hoog gaan.

Als we *odds* gebruiken als afhankelijke variabele,
krijgt onze voorspelling *van de kans* een ‘plafond’ van 1

# STAP 2: (NATUURLIJK) LOGARITME

Logit transformatie in twee stappen

# NATUURLIJK LOGARITME

$e^a = b$ $\quad$ $\ln b = a$ $\quad$ $e \approx 2{,}71828183$

$e^5 \approx 148{,}413$ $\quad$ $\ln 148{,}413 = 5$

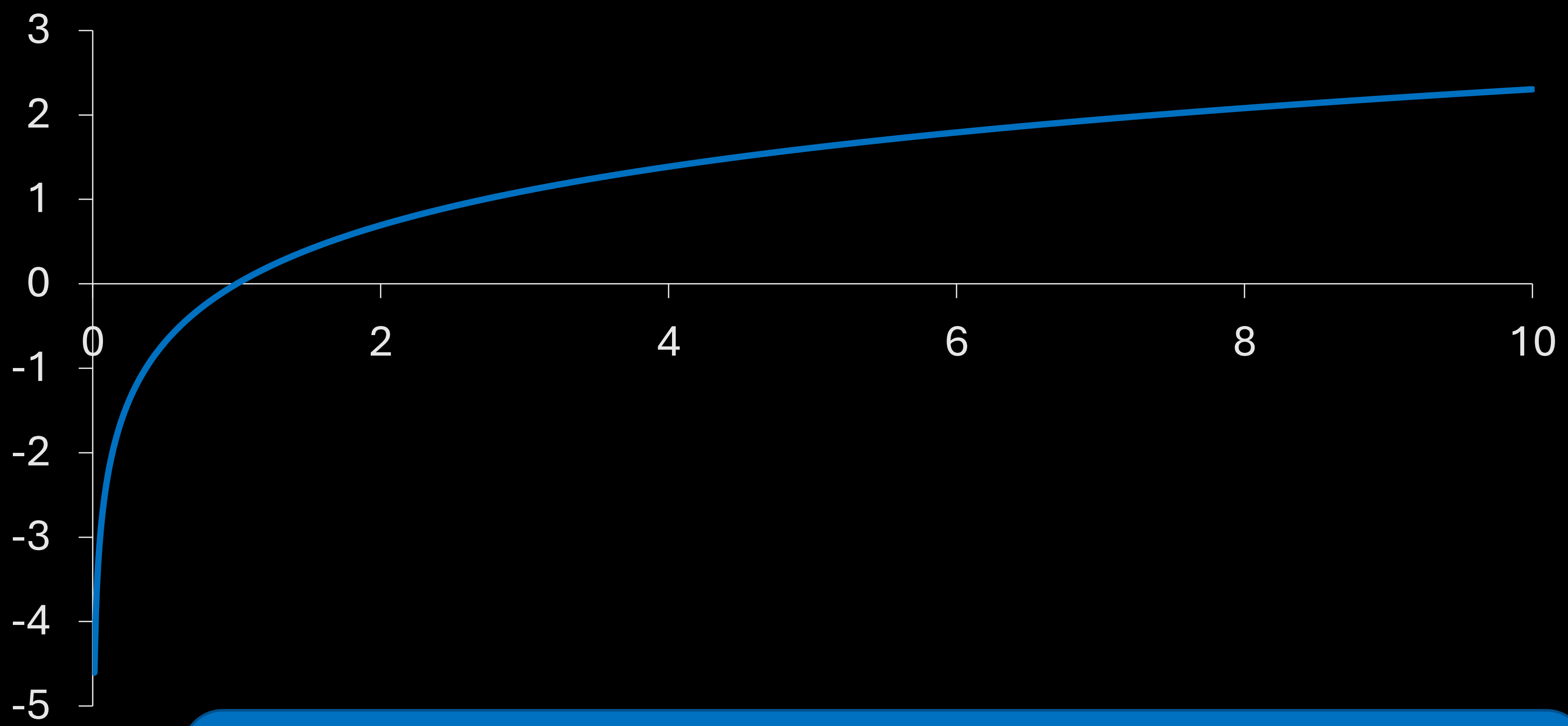
$e^1 = e$ $\quad$ $\ln e = 1$

$e^0 = 1$ $\quad$ $\ln 1 = 0$

$e^{-a} = \frac{1}{e^a}$ $\quad$ $\ln \frac{1}{e^a} = -a$

Er is geen logaritme voor negatieve getallen

# NATUURLIJK LOGARITME



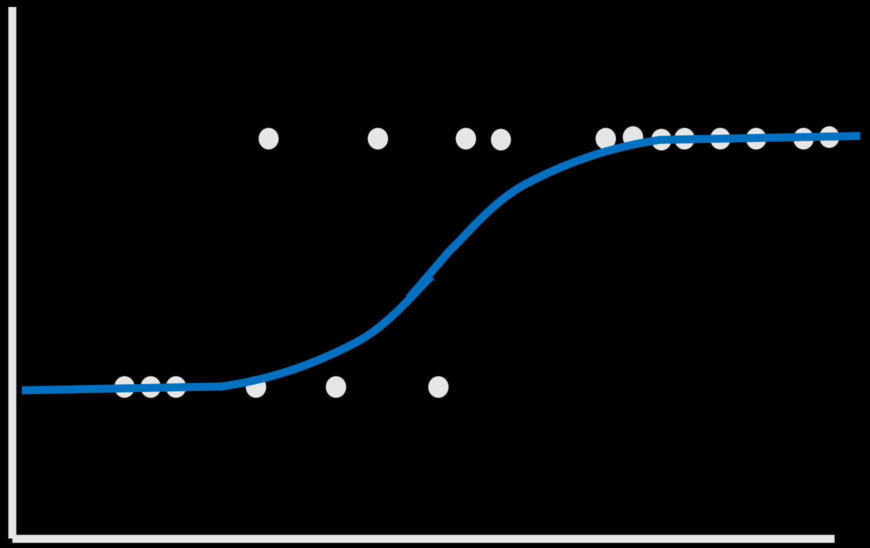

LN transformeert getallen van 0 tot +∞ in getallen van -∞ tot +∞

# LOGISTISCHE REGRESSIE

Regressie met logit transformatie

# LOGGED ODDS

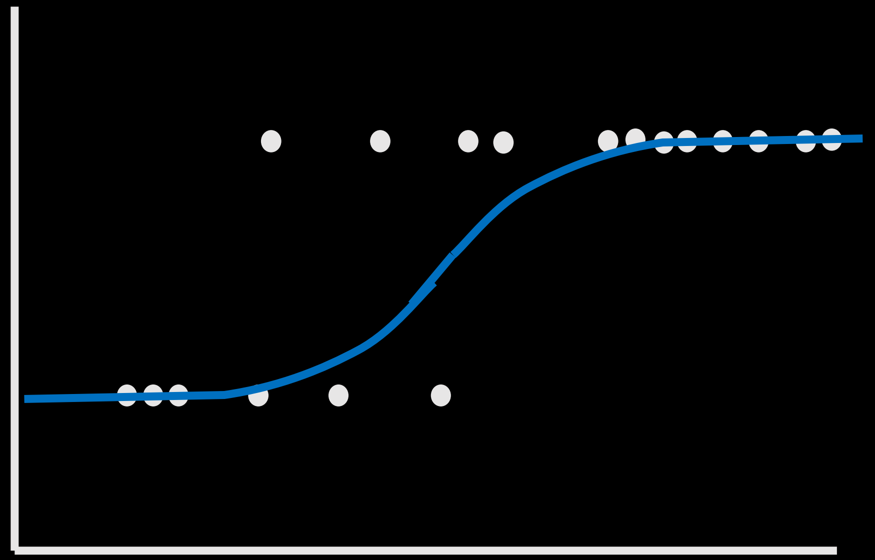
$$\ln \frac{P_i}{1 - P_i} = \beta_0 + \beta_1 X_i$$



1. Kans $P_i$: tussen 0 en 1 — plafond en vloer
2. Odds $\frac{P_i}{1-P_i}$: tussen 0 en +∞ — plafond weg
3. Logged odds $\ln \frac{P_i}{1-P_i}$: tussen -∞ en +∞ — plafond en vloer weg

# LOGGED ODDS

$$\ln \frac{P_i}{1 - P_i} = \beta_0 + \beta_1 X_i$$



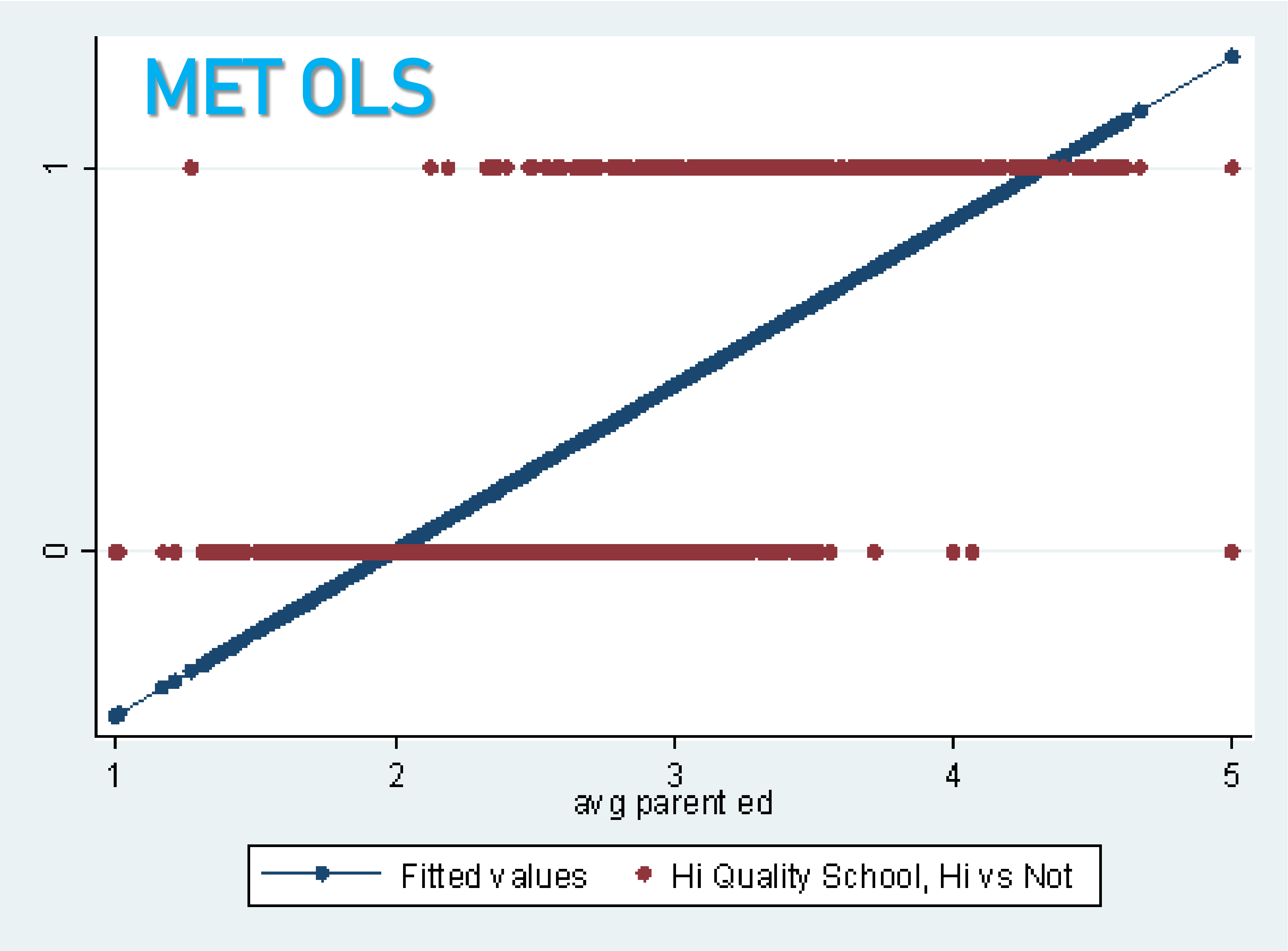
De *logged odds* kunnen voorspeld met een lineair model
Die kunnen we weer vertalen naar kansen tussen 0 en 1

https://elmarjansen.nl/os

# OEFENING 1

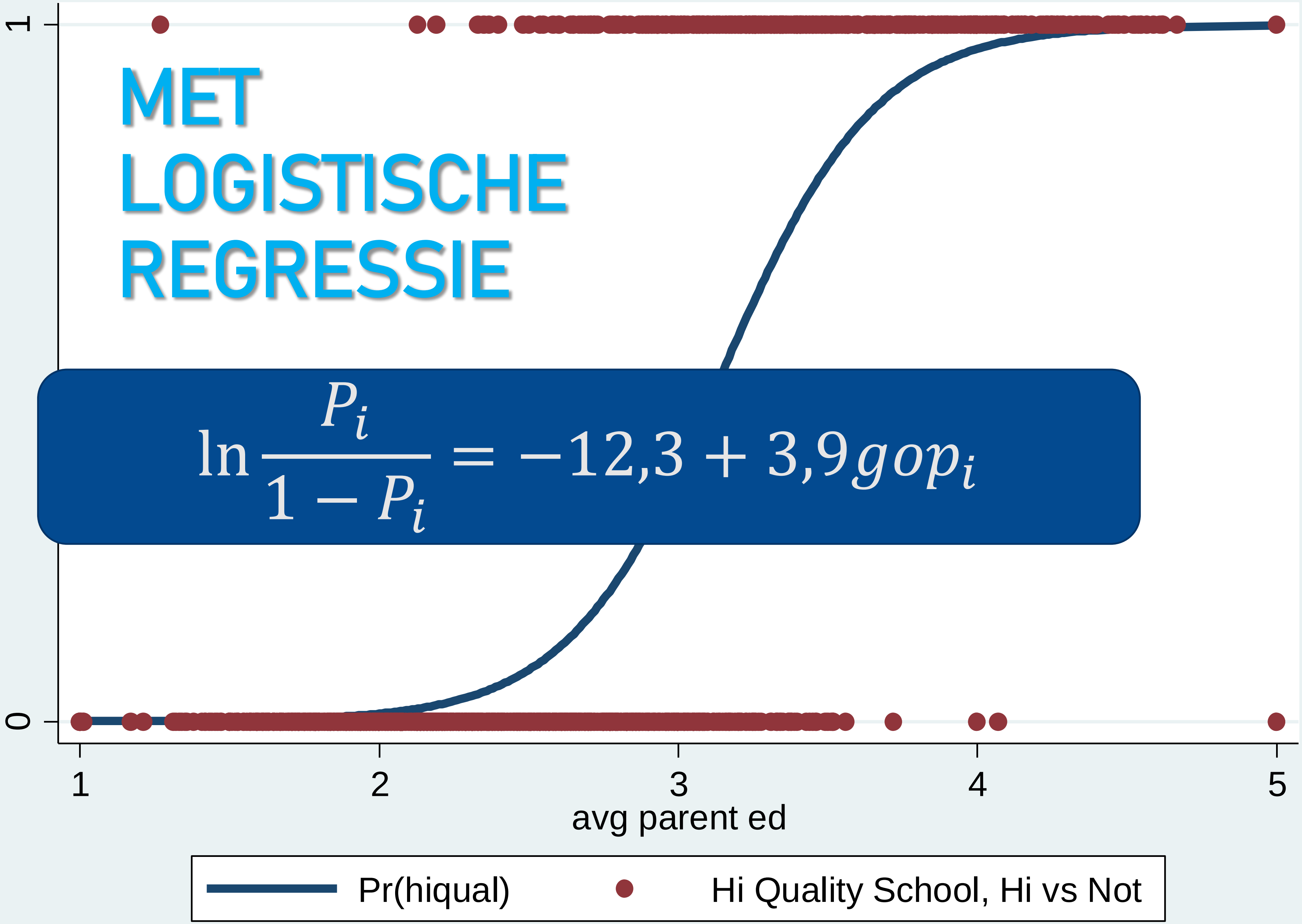
## Schoolkwaliteit in California

- 1158 Californische scholen
- Scholen scoren hoog of laag op kwaliteit
- We willen schoolkwaliteit voorspellen
  op basis van de opleiding van ouders (0-5)


MET OLS
1
0
1
2
3
4
5
avg parent ed
Fitted values
Hi Quality School, Hi vs Not

# MET LOGISTISCHE REGRESSIE

$$\ln \frac{P_i}{1 - P_i} = -12{,}3 + 3{,}9 gop_i$$



# IN R

```
> model <- glm(hiqual ~ avg_ed, data=df, family="binomial")
> summary(model)

Call:
glm(formula = hiqual ~ avg_ed, family = "binomial", data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3005     0.7315  -16.82   <2e-16 ***
avg_ed        3.9096     0.2383   16.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1461.37  on 1157  degrees of freedom
Residual deviance:  707.83  on 1156  degrees of freedom
  (42 observations deleted due to missingness)
AIC: 711.83

Number of Fisher Scoring iterations: 6
```

# IN R

```
> summary(model)

Call:
glm(formula = hiqual ~ avg_ed, family = "binomial", data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3005     0.7315  -16.82   <2e-16 ***
avg_ed        3.9096     0.2383   16.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1461.37  on 1157  degrees of freedom
Residual deviance:  707.83  on 1156  degrees of freedom
  (42 observations deleted due to missingness)
AIC: 711.83
```

$$\ln \frac{P_i}{1 - P_i} = \beta_0 + \beta_1 X_i$$

# IN R

```
> summary(model)

Call:
glm(formula = hiqual ~ avg_ed, family = "binomial", data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3005     0.7315  -16.82   <2e-16 ***
avg_ed        3.9096     0.2383   16.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

Dus: als opleiding van ouders 1 punt omhoog gaat, gaan de logged odds om een goede school te zijn omhoog met 3,9

$$\ln \frac{P_i}{1 - P_i} = -12,3 + 3,9 \times gop_i$$

IN R

# Hoe interpreteren we deze coëfficienten??

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -12.3005     0.7315  -16.82   <2e-16 ***
avg_ed        3.9096     0.2383   16.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

Dus: als opleiding van ouders 1 punt omhoog gaat, gaan de logged odds om een goede school te zijn omhoog met 3,9

$$\ln \frac{P_i}{1 - P_i} = -12{,}3 + 3{,}9 \times gop_i$$

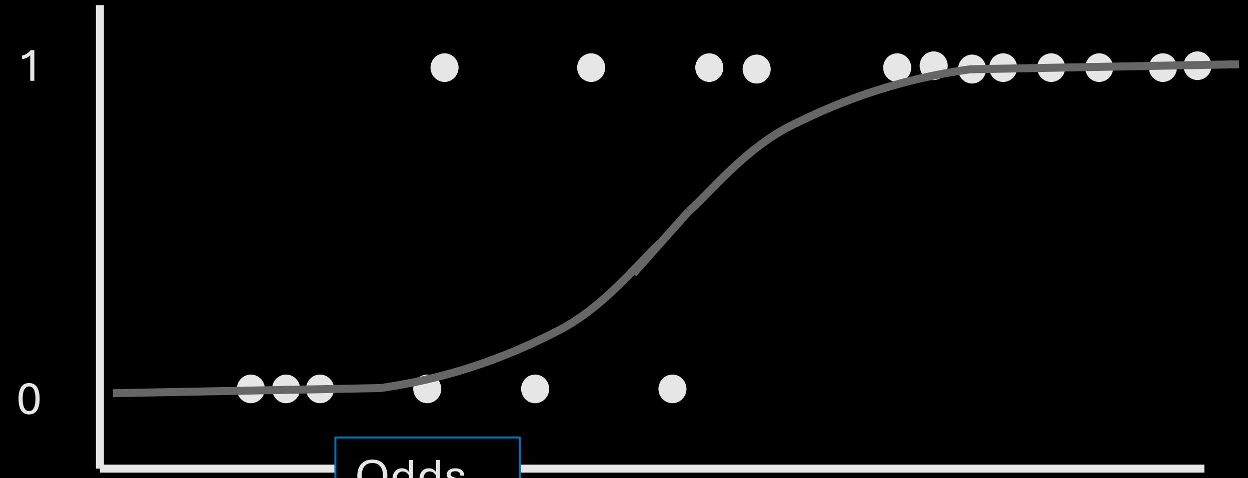
*Hoe interpreteren we deze coëfficienten??*


# HET LEZEN VAN DE RESULTATEN

Logistische regressie: wat?

# TOT NU TOE

Dichotome afhankelijke variabele



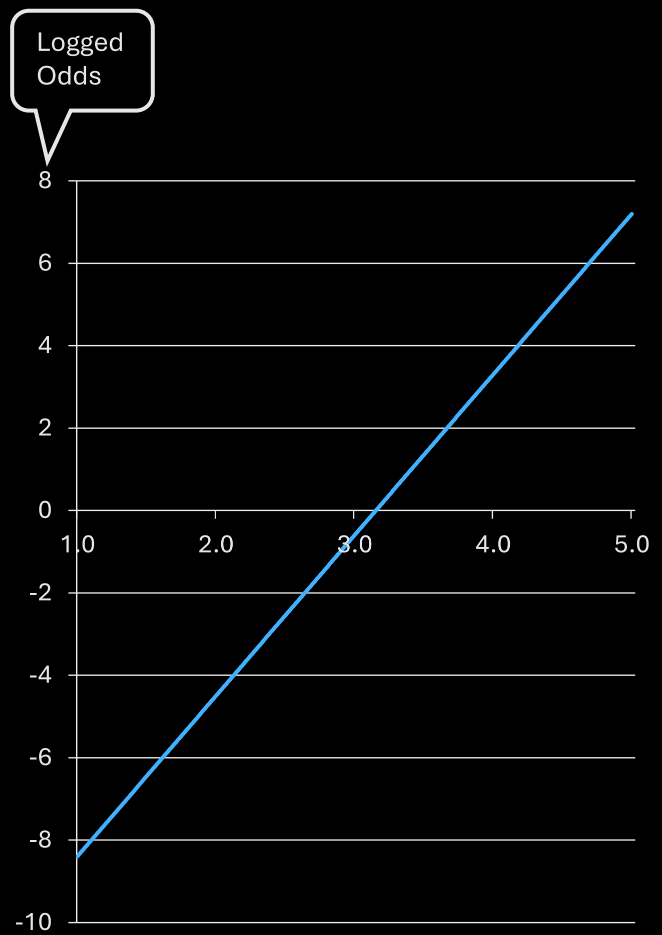

$$\underbrace{\ln \overbrace{\frac{P_i}{1-P_i}}^{\text{Odds}}}_{\text{Logged odds}} = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Lineair deel}}$$

# DE P UITPAKKEN: STAP 0

Logged Odds

Lineair

$$\ln \frac{P_i}{1 - P_i} = -12{,}3 + 3{,}9 \times gop_i$$



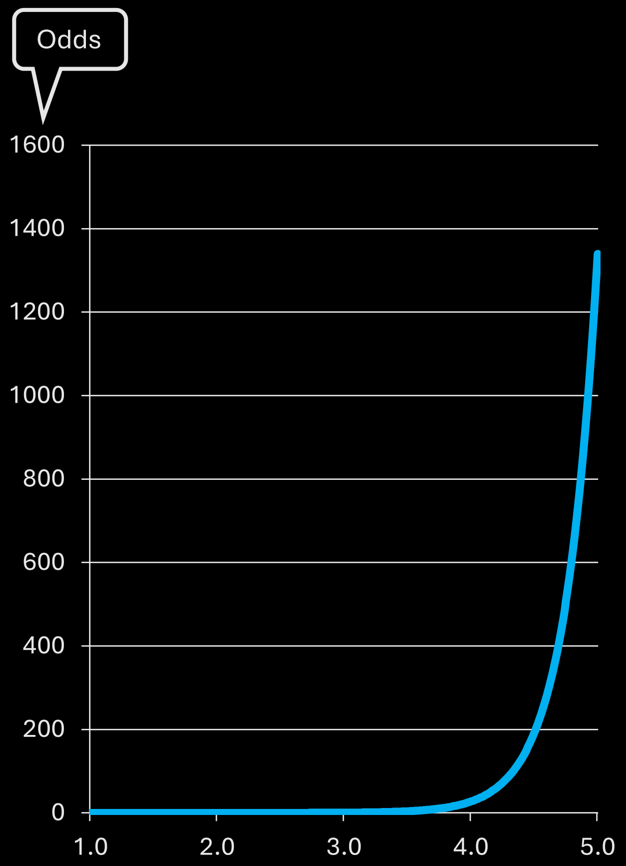
# DE P UITPAKKEN: STAP 1

$$\ln \frac{P_i}{1 - P_i} = -12{,}3 + 3{,}9 \times gop_i$$

Odds

Exponentieel

$$\frac{P_i}{1 - P_i} = e^{-12{,}3+3{,}9\times gop_i}$$



# DE P UITPAKKEN: STAP 2

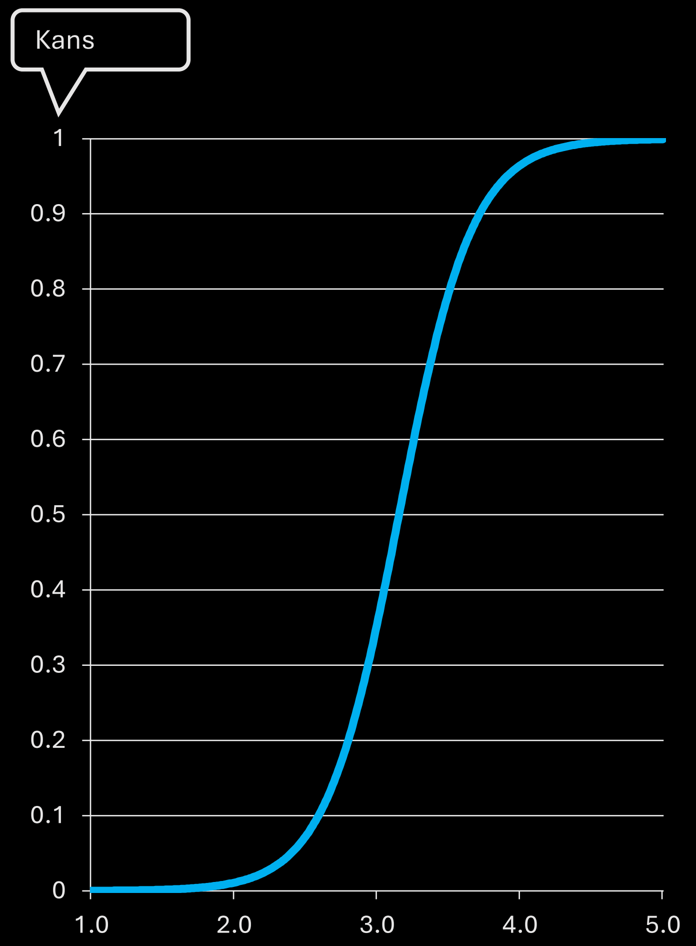
$$\ln \frac{P_i}{1 - P_i} = -12{,}3 + 3{,}9 \times gop_i$$

$$\frac{P_i}{1 - P_i} = e^{-12{,}3+3{,}9\times gop_i}$$

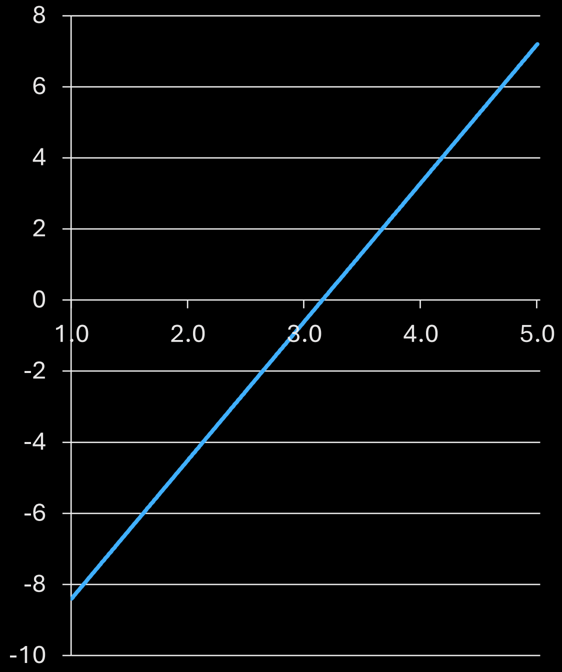
Probability

Logit

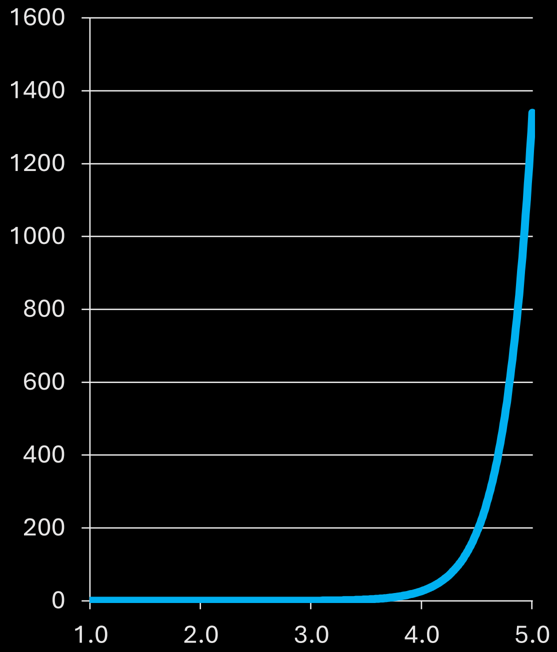
$$P_i = \frac{e^{-12{,}3+3{,}9\times gop_i}}{1 + e^{-12{,}3+3{,}9\times gop_i}}$$



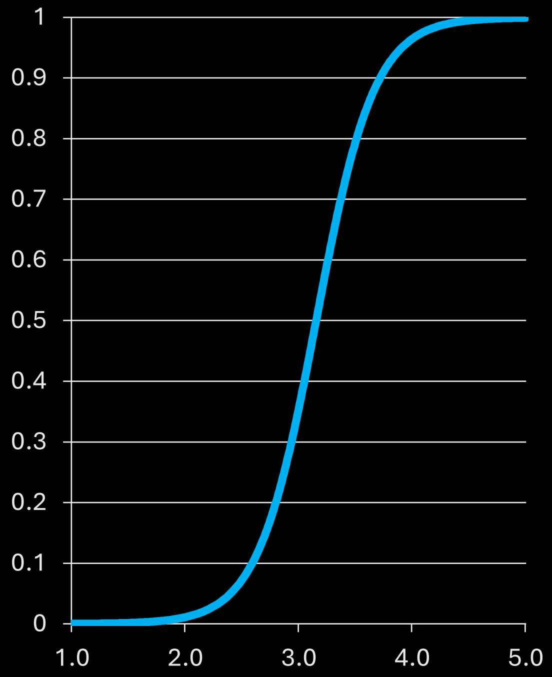
# LOGIT TRANSFORMATIE




$$\ln \frac{P_i}{1 - P_i} = -12{,}3 + 3{,}9 \times gop_i$$



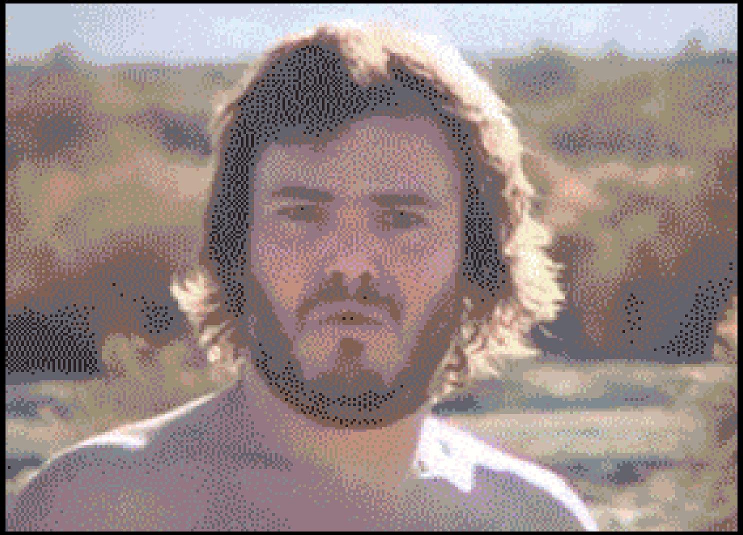

$$\frac{P_i}{1 - P_i} = e^{-12{,}3 + 3{,}9 \times gop_i}$$




$$P_i = \frac{e^{-12{,}3 + 3{,}9 \times gop_i}}{1 + e^{-12{,}3 + 3{,}9 \times gop_i}}$$

# DE PUZZEL

Wat moeten we met deze coëfficienten?



$$\ln \frac{P_i}{1 - P_i} = \beta_0 + \beta_1 X_i$$

# INTERPRETEREN : 3 MANIEREN

1. effect op de logged odds:

$$\ln \frac{P_i}{1 - P_i} = \beta_0 + \beta_1 X_i$$

2. effect op de odds:

$$\frac{P_i}{1 - P_i} = e^{\beta_0 + \beta_1 X_i}$$

3. effect op de voorspelde kansen:

$$P_i = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$$

# 1. EFFECT OP DE LOGGED ODDS

Effecten interpreteren

# 1. EFFECT OP DE LOGGED ODDS

# 1. EFFECT OP DE LOGGED ODDS

**Table 3**
**Policy outcomes and the policy preferences of average citizens, economic elites, and interest groups**

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| Preferences of average citizens | .64 (.08)*** | — | — | .03 (.08) |
| Preferences of economic elites | — | .81 (.08)*** | — | .76 (.08)*** |
| Alignment of interest groups | — | — | .59 (.09)*** | .56 (.09)*** |
| R-sq | .031 | .049 | .028 | .074 |

***p<.001

Note: All predictors are scaled to range from 0 to 1. The dependent variable is the policy outcome, coded 1 if the proposed policy change took place within four years of the survey date and 0 if it did not. Predictors are the logits of the imputed percent of respondents at the fiftieth ("average citizens") or ninetieth ("economic elites") income percentile that favor the proposed policy change, and the Net Interest-Group Alignment Index described in the text. Standard errors are asymptotically distribution-free, and all analyses reflect estimated measurement error in the predictors, as described in Appendix 2. The standardized coefficients for model 4 in this table are .01, .21, and .16 for average citizens, economic elites, and interest groups, respectively. N=1,779.

$$\ln \frac{P_i}{1 - P_i} = \beta_0 + \boxed{\beta_1} PrevAv_i + \beta_2 PrevRich_i + \beta_3 IntrGr_i$$

# 1. EFFECT OP DE LOGGED ODDS

**Table 3**
**Policy outcomes and the policy preferences of average citizens, economic elites, and interest groups**

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| Preferences of average citizens | .64 (.08)*** | — | — | .03 (.08) |
| Preferences of economic elites | — | .81 (.08)*** | — | .76 (.08)*** |
| Alignment of interest groups | — | — | .59 (.09)*** | .56 (.09)*** |
| R-sq | .031 | .049 | .028 | .074 |

***p<.001

Note: All predictors are scaled to range from 0 to 1. The dependent variable is the policy outcome, coded 1 if the proposed policy change took place within four years of the survey date and 0 if it did not. Predictors are the logits of the imputed percent of respondents at the fiftieth ("average citizens") or ninetieth ("economic elites") income percentile that favor the proposed policy change, and the Net Interest-Group Alignment Index described in the text. Standard errors are asymptotically distribution-free, and all analyses reflect estimated measurement error in the predictors, as described in Appendix 2. The standardized coefficients for model 4 in this table are .01, .21, and .16 for average citizens, economic elites, and interest groups, respectively. N=1,779.

$$\ln \frac{P_i}{1 - P_i} = \beta_0 + \boxed{.03} PrevAv_i + .76 PrevRich_i + .56 IntrGr_i$$

Interpretatie effect ($\beta_1$) op logged odds:

Voor iedere eenheid toename van PrevAv, gaan de logged odds (dat een plan wordt aangenomen) omhoog met 0.03.

# 1. EFFECT OP DE LOGGED ODDS

**Table 3**
**Policy outcomes and the policy preferences of average citizens, economic elites, and interest groups**

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| Preferences of average citizens | .64 (.08)*** | — | — | .03 (.08) |
| Preferences of economic elites | — | .81 (.08)*** | — | .76 (.08)*** |
| Alignment of interest groups | — | — | .59 (.09)*** | .56 (.09)*** |
| R-sq | .031 | .049 | .028 | .074 |

***p<.001

Note: All predictors are scaled to range from 0 to 1. The depen coded 1 if the propose years of the survey date the logits of the imputed percent of respondents at the fiftieth ("average citizens") c rcentile that favor t nterest-Group Alig d errors are asymp reflect estimated measurement error in the predictors, as described in Appendix 2. The standardized coefficients for model 4 in this table are .01, .21, and .16 for average citizens, economic elites, and interest groups, respectively. N=1,779.

Het effect is positief

Het effect verschilt niet significant van 0

<u>Interpretatie effect ($\beta_1$) op logged odds:</u>

Voor iedere eenheid toename van PrevAv, gaan de logged odds (dat een plan wordt aangenomen) omhoog met 0.03.

# 2. EFFECT OP DE ODDS

Effecten interpreteren

# 2. EFFECT OP DE ODDS

$$\ln \frac{P_i}{1-P_i} = \beta_0 + \beta_1 X_i$$

$$e^a = b$$
$$\ln b = a$$

$$\frac{P_i}{1-P_i} = e^{\beta_0+\beta_1 X_i}$$

$$a^{b+c} =$$
$$a^b * a^c$$

$$\frac{P_i}{1-P_i} = \left(e^{\beta_0}\right) \times \left(e^{\beta_1}\right)^{X_i}$$

# 2. EFFECT OP DE ODDS

Wat gebeurt er met de odds als X stijgt met 1?

$$\frac{P_i}{1 - P_i} = (e^{\beta_0}) \times (e^{\beta_1})^{X_i}$$

# 2. EFFECT OP DE ODDS

Wat gebeurt er met de odds als X stijgt met 1?

$$\frac{P_i}{1 - P_i} = e^{\beta_0} \times e^{\beta_1 (X_i + 1)}$$

# 2. EFFECT OP DE ODDS

$a^{b+c} = a^b * a^c$

Wat gebeurt er met de odds als X stijgt met 1?

$$\frac{P_i}{1 - P_i} = e^{\beta_0} \times e^{\beta_1 (X_i + 1)}$$

$$= e^{\beta_0} \times e^{\beta_1 X_i + \beta_1}$$

$$= e^{\beta_0} \times e^{\beta_1 X_i} \times e^{\beta_1}$$

# 2. EFFECT OP DE ODDS

Wat gebeurt er met de odds als X stijgt met 1?

$$\frac{P_i}{1-P_i} = e^{\beta_0} \times e^{\beta_1 X_i} \times e^{\beta_1}$$

Als X met 1 stijgt, worden de odds *vermenigvuldigd* met $e^{\beta_1}$

# 2. EFFECT OP DE ODDS

**Table 3**
**Policy outcomes and the policy preferences of average citizens, economic elites, and interest groups**

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| Preferences of average citizens | .64 (.08)*** | — | — | .03 (.08) |
| Preferences of economic elites | — | .81 (.08)*** | — | .76 (.08)*** |
| Alignment of interest groups | — | — | .59 (.09)*** | .56 (.09)*** |
| R-sq | .031 | .049 | .028 | .074 |

***p<.001

Note: All predictors are scaled to range from 0 to 1. The dependent variable is the policy outcome, coded 1 if the proposed policy change took place within four years of the survey date and 0 if it did not. Predictors are the logits of the imputed percent of respondents at the fiftieth ("average citizens") or ninetieth ("economic elites") income percentile that favor the proposed policy change, and the Net Interest-Group Alignment Index described in the text. Standard errors are asymptotically distribution-free, and all analyses reflect estimated measurement error in the predictors, as described in Appendix 2. The standardized coefficients for model 4 in this table are .01, .21, and .16 for average citizens, economic elites, and interest groups, respectively. N=1,779.

$$\frac{P_i}{1-P_i} = e^{\beta_0} \times \boxed{e^{.03 PrevAv_i}} \times e^{.76 PrevRich_i} \times e^{.56 Inter_i}$$

Als PrevAv = 0

$e^{0.03\times 0} = e^{0} = 1$

Als PrevAv = 1

$e^{0.03\times 1} = e^{0.03}$

Als PrevAv = 2

$e^{0.03\times 2} = e^{0.03} \times e^{0.03}$

Als PrevAv = 3

$e^{0.03\times 3} = e^{0.03} \times e^{0.03} \times e^{0.03}$

# 2. EFFECT OP DE ODDS

**Table 3**
**Policy outcomes and the policy preferences of average citizens, economic elites, and interest groups**

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| Preferences of average citizens | .64 (.08)*** | — | — | .03 (.08) |
| Preferences of economic elites | — | .81 (.08)*** | — | .76 (.08)*** |
| Alignment of interest groups | — | — | .59 (.09)*** | .56 (.09)*** |
| R-sq | .031 | .049 | .028 | .074 |

***p<.001

Note: All predictors are scaled to range from 0 to 1. The dependent variable is the policy outcome, coded 1 if the proposed policy change took place within four years of the survey date and 0 if it did not. Predictors are the logits of the imputed percent of respondents at the fiftieth ("average citizens") or ninetieth ("economic elites") income percentile that favor the proposed policy change, and the Net Interest-Group Alignment Index described in the text. Standard errors are asymptotically distribution-free, and all analyses reflect estimated measurement error in the predictors, as described in Appendix 2. The standardized coefficients for model 4 in this table are .01, .21, and .16 for average citizens, economic elites, and interest groups, respectively. N=1,779.

$$\frac{P_i}{1-P_i} = e^{\beta_0} \times \boxed{e^{.03PrevAv_i}} \times e^{.76PrevRich_i} \times e^{.56IntGr_i}$$

If PrevAv = 0
$e^{0.03*0} = e^{0} = 1$

If PrevAv = 1
$e^{0.03*1} = e^{0.03}$

Als PrevAv stijgt met 1, worden de ***odds vermenigvuldigd*** met $e^{0.03}$ of ongeveer 1.0304545.

# 2. EFFECT OP DE ODDS

$$\frac{P_i}{1-P_i} = e^{\beta_0} \times e^{.03PrevAv_i} \times e^{.76PrevRich_i} \times e^{.56IntrGr_i}$$

Als PrevAv stijgt met 1, worden de ***odds vermenigvuldigd*** met $e^{0.03}$ of ongeveer 1.0304545.

$e^{\beta_1}$ *noemen we de* ***odds ratio***

# ODDS RATIO

$$OR_1 = e^{\beta_1}$$
$$OR_2 = e^{\beta_2}$$

etc.

**Interpretatie:**

Als X stijgt, worden de *odds op y* vermenigvuldigd met de OR

0 < oddsratio < 1

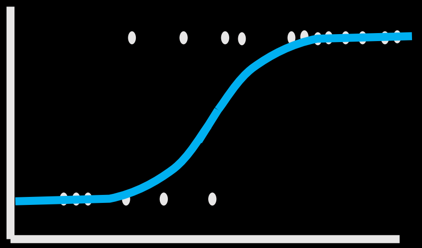
odds dalen
als X stijgt

oddsratio > 1:

odds stijgen
als X stijgt

# 3. EFFECTEN OP DE VOORSPELDE KANSEN

# 3. EFFECTEN OP DE VOORSPELDE KANSEN



$$odds_i = \frac{P_i}{1 - P_i} = e^{\beta_0 + \beta_1 X_i} = e^{\beta_0} e^{\beta_1 X_i}$$

$$P_i = \frac{odds_i}{1 + odds_i} = \frac{e^{\beta_0} e^{\beta_1 X_i}}{1 + e^{\beta_0} e^{\beta_1 X_i}}$$

Je kunt kansen alleen berekenen voor *specifieke waarden* van je onafhankelijke variabele

# 3. EFFECTEN OP DE VOORSPELDE KANSEN

Beginnen met de coefficienten

```
> model <- glm(adoptedin4 ~ PRED50_SW + PRED90_SW, data=DS3, family="binomial")
> coef(model)
(Intercept)   PRED50_SW   PRED90_SW 
  -1.957883   -1.976131    4.204488 
```

# 3. EFFECTEN OP VOORSPELDE KANSEN

We kunnen R ook vragen de odds ratios te berekenen met de exp-functie:

Als voorstel gaat van geen steun bij elite (0) naar volledige steun van elite (1), worden de *odds* dat een voorstel wordt aangenomen 67 keer zo groot

```
> model <- glm(adoptedin4 ~ PRED50_SW + PRED90_SW, data=DS3, family="binomial")
> coef(model)
(Intercept)    PRED50_SW    PRED90_SW
  -1.957883    -1.976131     4.204488
> exp(coef(model))
(Intercept)    PRED50_SW    PRED90_SW
  0.1411569    0.1386044   66.9862940
```

# 3. EFFECTEN OP VOORSPELDE KANSEN

```
> model <- glm(adoptedin4 ~ PRED50_SW + PRED90_SW, data=DS3, family="binomial")
> coef(model)
(Intercept)   PRED50_SW   PRED90_SW
  -1.957883   -1.976131    4.204488
> exp(coef(model))
(Intercept)   PRED50_SW   PRED90_SW
  0.1411569   0.1386044  66.9862940
```

*Bijvoorbeeld: kans op voorstel dat alle modale burger willen, maar de rijkste 10 procent niet:*

$$odds_i = 0{,}14 \times 0{,}14^{x1} \times 67^{x2}$$
$$= 0{,}14 \times 0{,}14^{1} \times 67^{0}$$
$$= 0{,}14 \times 0{,}14 \times 1 = 0{,}019$$

$$P_i = \frac{0{,}019}{1 + 0{,}019} = 0{,}019$$

# 3. EFFECTEN OP VOORSPELDE KANSEN

```
> model <- glm(adoptedin4 ~ PRED50_SW + PRED90_SW, data=DS3, family="binomial")
> coef(model)
(Intercept)   PRED50_SW   PRED90_SW
 -1.957883   -1.976131    4.204488
> exp(coef(model))
(Intercept)   PRED50_SW   PRED90_SW
 0.1411569   0.1386044  66.9862940
```

*Bijvoorbeeld: kans op voorstel dat geen enkele modale burger wil, maar de volledige rijkste 10 procent wel:*

$$odds_i = 0{,}14 \times 0{,}14^{x1} \times 67^{x2}$$
$$= 0{,}14 \times 0{,}14^{0} \times 67^{1}$$
$$= 0{,}14 \times 1 \times 67 = 9{,}45$$

$$P_i = \frac{9{,}45}{1 + 9{,}45} = 0{,}90$$

# HOE INTERPRETEREN WE DE RESULTATEN?

Logistische regressie: wat???

# INTERPRETEREN : 3 MANIEREN

1. op de logged odds

$$\ln \frac{P_i}{1 - P_i} = \beta_0 + \beta_1 X_i$$

Voor iedere stijging van 1 van X, stijgen de *logged odds* met $\beta$.

2. op de odds

$$\frac{P_i}{1 - P_i} = e^{\beta_0 + \beta_1 X_i}$$

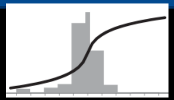
Voor iedere stijging van 1 van X, worden de *odds vermenigvuldigd* met $e^{\beta}$.

3. op de voorspelde kansen

$$P_i = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$$

“Als X=a, is de *kans* .., als X=b, is de *kans* …”

Of grafiek



# INTERPRETEREN : 3 MANIEREN

| ***Coëfficiënt logged odds*** | ***Odds ratios*** |
|---|---|
| < 0:<br>logged odds dalen als X stijgt | < 1:<br>odds dalen als X stijgt |
| Optellen | Vermenigvuldigen |
| > 0:<br>logged odds stijgen als X stijgt | > 1:<br>odds stijgen als X stijgt |

# WHAT IS 'PSEUDO R2'?

McFadden Pseudo R2: Hoe goed is dit model?

# EEN $R^2$ VOOR LOGISTISCHE REGRESSIE?

Met logistische regressie voorspellen we *kansen*, dus er is geen *residu*


Je kunt dus ook niet spreken van *verklaard variantie*

OLS:

$$Y_i = b_0 + b_1 X_i + e_i$$

"verklaarde variantie" voor logistische regressie?

Logistisch:

$$\ln \frac{P_i}{1 - P_i} = b_0 + b_1 x_i$$

# EEN $R^2$ FOR LOGISTISCHE REGRESSIE?

OLS:

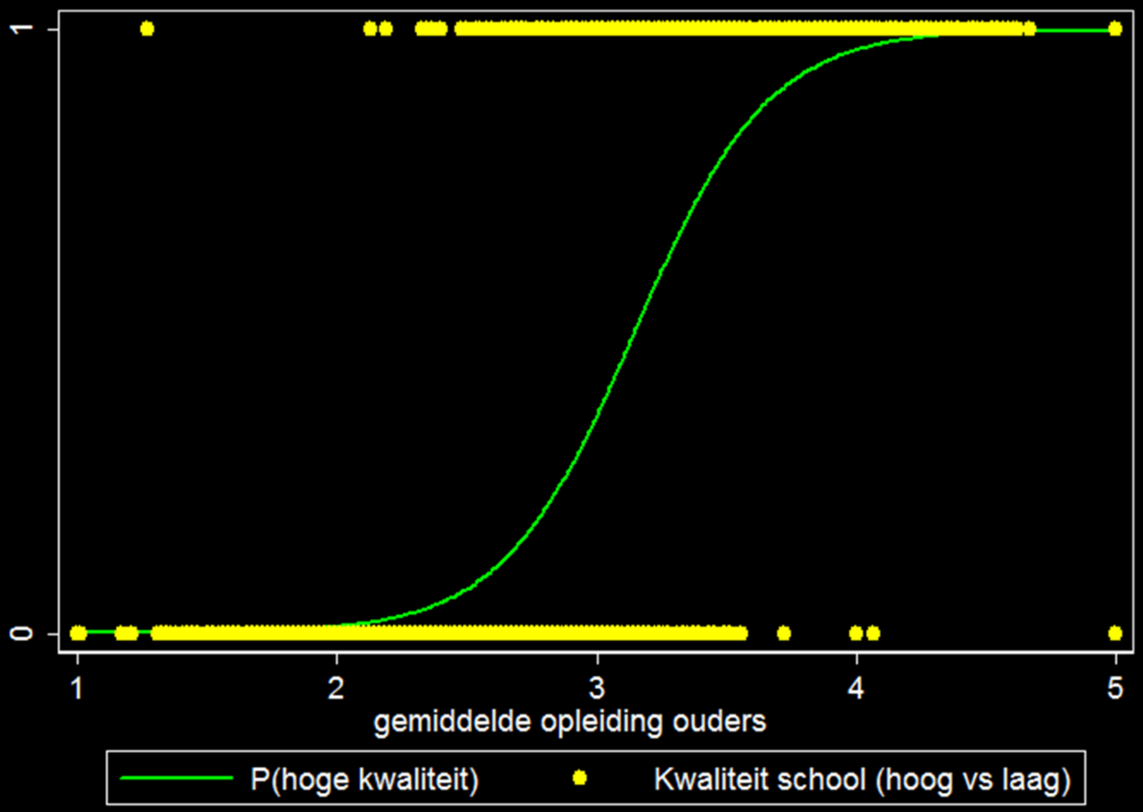
$$Y_i = b_0 + b_1 X_i + e_i$$

Logistisch:

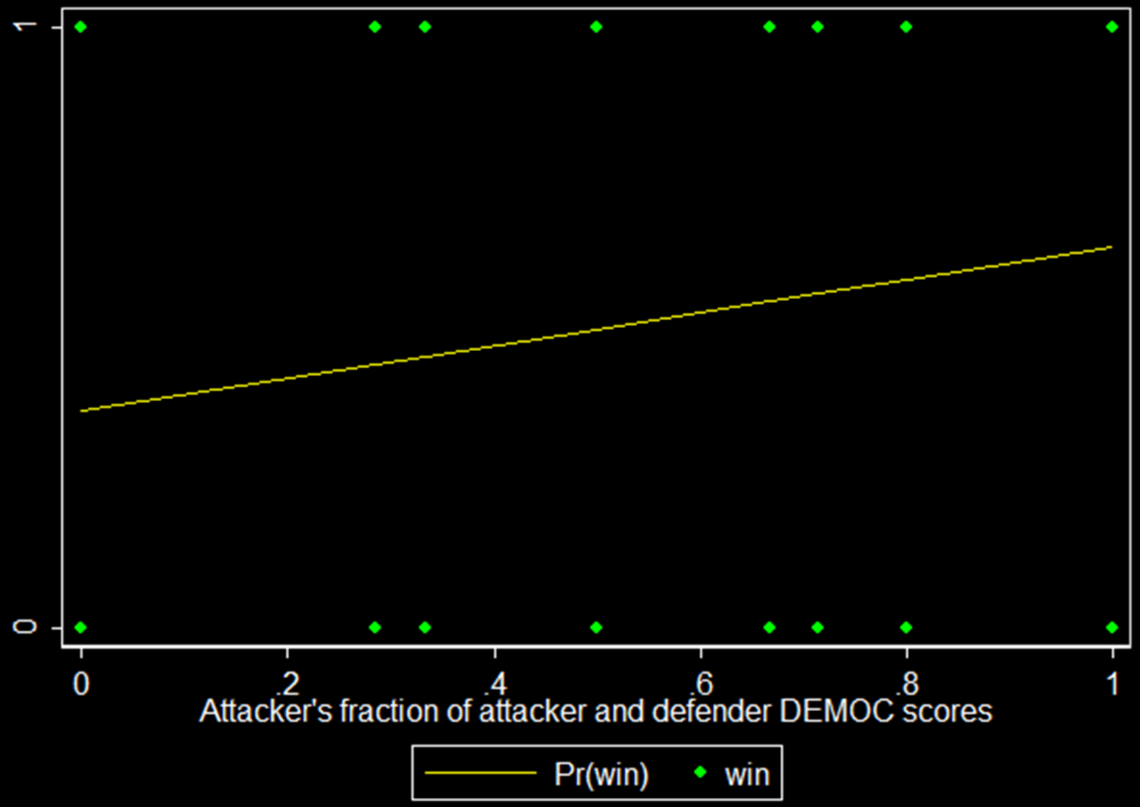
$$\ln \frac{P_i}{1 - P_i} = b_0 + b_1 x_i$$

Geen residu => geen “verklaarde variantie”

# WANNEER ZOU JE EEN HOGE "$R^2$" VERWACHTEN?



# WANNEER ZOU JE EEN LAGE "$R^2$" VERWACHTEN?



# MCFADDEN PSEUDO $R^2$

Vergelijkt de likelihood van het gebruikte model, met dat van het model met 0 coëfficienten (het ‘nulmodel’)

$$\frac{\underbrace{(-2\ln L_0)}_{\text{Likelihood van deze steekproef, als \textbf{alle coefficients 0 zijn}}} - \underbrace{(-2\ln L_1)}_{\text{Likelihood van deze steekproef, als \textbf{dit model correct is}}}}{(-2\ln L_0)}$$

*Likelihood (of de log-likelihood) is een maat voor hoe goed onze data bij het model past (zie verderop voor details)*

# MCFADDEN PSEUDO R²

Log Likelihood Ratio: toename van likelihood die we aan ons model kunnen toegschrijven

$$\frac{(-2\ln L_0) - (-2\ln L_1)}{(-2\ln L_0)}$$

In verhouding tot de likelihood als er geen model is (coefficienten 0)

# MCFADDEN PSEUDO $R^2$

Net als met normale $R^2$:

1 is een perfect voorspellend model

0 is een model zonder enige voorspellende waarde

$$\frac{(-2\ln L_0) - (-2\ln L_1)}{(-2\ln L_0)}$$

Bedenk dat waarden vaak wat lager zijn dan van normale $R^2$ en dat het geen “verklaarde variantie” meer is

# PSEUDO R2 BEREKENEN IN R

$(-2 \ln L_0) \approx 1461.37416$

```
Call:
glm(formula = hiqual ~ avg_ed, family = "binomial", data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3005     0.7315  -16.82   <2e-16 ***
avg_ed        3.9096     0.2383   16.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1461.37  on 1157  degrees of freedom
Residual deviance:  707.83  on 1156  degrees of freedom
  (42 observations deleted due to missingness)
AIC: 711.83

Number of Fisher Scoring iterations: 6
```

$(-2 \ln L_1) \approx 707.83438$

$$\frac{1461 - 708}{1461}$$

$$\frac{(-2 \ln L_0) - (-2 \ln L_1)}{(-2 \ln L_0)}$$

$$\frac{(-2\ln L_0) - (-2\ln L_1)}{(-2\ln L_0)} \qquad \frac{1461.37416 - 707.83438}{1461.37416}$$

$$\approx 0.51563781584$$

```
> # Pseudo R2 berekenen:
> with(summary(model), 1 - deviance/null.deviance)
[1] 0.5156378
```

# PSEUDO R2 IN SPSS

SPSS geeft twee andere bekende “Pseudo $R^2$”-maten:

- Nagelkerke Pseudo $R^2$
- Cox & Snell Pseudo $R^2$

Bedenk ook hier: het gaat dus niet om een echte $R^2$ (want geen verklaarde variantie)

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 20.856[a] | .505 | .676 |

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

# VOLGENDE WEEK

Multilevel Analyse

# LOGISTISCHE REGRESSIE

4 juni 2024

Training O + S

Elmar Jansen (elmar@elmarjansen.nl)